Data Mining E-learning Data for a Student Prediction System

Fahad Alghamdi

Supervisors: Dr. Richard Jensen
Prof. Qiang Shen

Ph.D. Thesis
Department of Computer Science
Institute of Mathematics, Physics and Computer Science
Aberystwyth University

March 16, 2017

Declaration and Statement

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated.

Where **correction services**[1] have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... (candidate)

Date ....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

[1] This refers to the extent to which the text has been corrected by others.

Abstract

A large body of literature has been published regarding E-learning and improving students' performance through the use of specialised advising systems. However, much less research has focussed on the use of data mining for E-learning systems to develop an effective performance prediction and advising tool. This study contributes to the literature in these under-researched areas.

Predicting students' performance in E-learning is not dependable on traditional face-to-face education methods, where the instructor interacts and receives direct feedback from their students. In E-learning systems, student performance prediction is based on analysing students' marks and progress. This thesis aims to apply data mining to all available data in the E-learning environment so that the process and effectiveness can be improved by predicting students' performance and suggesting advice to both the instructors and students based on the predictions made. The study thus aims to use data mining in order to improve the accuracy of previous prediction models, by gaining additional insights from student data. Details of previous prediction models and how these have been used or replaced by the new proposed system will also be presented.

This study adopts an experimental research approach where the researcher developed and tested an advising e learning software on a sample of students at the University of Dammam. The study has resulted in the production of a software that is able to give students advice based on their unique virtual learning experience, analysing metrics such as number and type of activities undertaken in the learning environment. The advice consists of suggestions regarding the best way to achieve higher predicted results, offering students clear, practical and quantifiable solutions (clearly measurable and having well-set rules and proposed solutions) to achieving higher academic performance. For instance, suggestions can include taking a number of additional quizzes or spending a certain time on a particular exercise. The software was tested on a sample of students and has resulted in the average scores of students significantly increasing by 5% from 67% to 72%. This suggest that the software can be used to improve students' performance if applied to larger samples.

Acknowledgements

I would like primarily to praise and thank ALLAH, the most Gracious and the most Merciful, who has given me the ability to do this work. Although few words do not do justice their contribution and support, I would like to thank many people for making this work possible. In particular, I would like to express my deepest gratitude and sincere thanks to my supervisor Dr Richard Jensen for his encouragement, patience, support and valuable advice; and my supervisor professor Qiang Shen for his patience, interest, support, and unlimited intellectual advice which has been very inspirational. I extend my thanks to examiners, Professor Trevor Male and Professor David Ellis. Special gratitude to my beloved country "Saudi Arabia" which gave me the opportunity to conduct my research in the UK by sponsoring this research.

I offer gratitude and thanks to the entire academic staff at the Faculty of Education at Dammam University who participated in this study as without their help, this study would not have been completed.

Also, I would like to thank all staff members of the Saudi Educational and Culture Bureau in London for their guidance, support and patience during this long journey. I would like to present my sincere thanks and appreciation to computer science department staff members and my colleagues for their co-operation and support. I feel very lucky to have undertaken this research at Aberystwyth University. I have been supported and encouraged by everyone, which made me feel that I am in my home. Last, and by no means least, I am very grateful to my extended and immediate family for their support. I would like especially to thank my dear wife "Reem" for her unlimited love, support and encouragement, my son "Osama" who have been my source of inspiration, my daughter "Shaden", my father, mother, brothers and sisters for their unlimited encouragement and support.

Table of Contents

List of Figures

List of tables

List of Abbreviations

AHS	Adaptive Hypermedia Systems .
AIWBES	Adaptive and Intelligent Web-Based Educational Systems.
UOD	University of Dammam
CART	Classification And Regression Trees.
CHAID	CHi-squared Automatic Interaction Detection .
DDL	Deanship of Distance Learning.
DM	Data Mining.
ML	Machine Learning
DT	Decision Tree.
EMES	e-Learning Management Electronic System.
HTML	Hypertext Markup Language.
ITS	Intelligent Tutoring Systems .
KFU	King Fahad University .
KDD	Knowledge Discovery.
LCMS	Learning Content Management Systems.
SPS	Students Prediction System.
SQL	Structured Query Language.
WWW	World Wide Web.
UI	User Interface

Chapter 1 Introduction

1.1 Introduction

The continuous rapid growth of the World Wide Web with its enormous array of browsing tools and resources have made it an incredibly important platform for collecting, sharing and distributing information. Most organizations and companies in existence today rely on the web for communication, marketing and even trading, making the WWW an ordinary, everyday tool, integral to the lives of the majority of people.

Unsurprisingly the web is the communication medium of choice for advanced, modern education systems where distance learning fast became the most widespread and efficient method of delivering course material, studying and sharing information.

The term ‘E-learning’ has now become an accepted standard for web-based learning, based on using computer networks to create, deliver, manage and support on-line education with a huge potential to improve online education and by making education available to a wider audience who might otherwise not be able to participate in traditional classroom learning. Furthermore, E-learning can contribute to traditional classroom learning by offering extra functionality, interactive learning, enhanced research facilities and also by enhancing the administration functions of a school and the way teachers and students communicate and organise themselves. Additionally, E-learning systems offer various methods of learning to suit different people, for example, self-paced, collaborative or tutorial, all within a common application.

Some common E-learning systems such as Virtual-U and Web-CT include delivery tools for course content, synchronous and asynchronous conference calling, survey, polling and quiz modules, virtual spaces for sharing and exchanging information, white boards, grade reporting systems, logbooks, assignment submission components, all making up efficient and easy to use management platforms for the general learning experience.

Such systems can also provide a substantial management aid for students by keeping records of what work was done, reading, writing, tasks, tests, assignments completed and those still to be done, and can even assist with communication within a particular course, particularly useful for distance learning courses.

Personal information can be safely stored within an E-learning system (a user profile) and as a result, E-learning systems accumulate substantial amounts of information, which when used for analysis become an invaluable source of educational data. It has to be said, however, that to manage successfully and analyse the gathered data is a complex and time-consuming process if done manually. For this reason, both instructors and students are in need of tools to help in various activities; instructors for assisting in analysing this data and tracking student activities and course structure and content, and students for advice on how to improve performance and get the most out of the system. This is when the virtues of data mining become apparent.

The use of data mining in E-learning is still in the early stages, but much research has been carried out and put into practice in certain areas of e-commerce where customer behaviour relating to purchases and satisfaction is fast becoming accepted practice. Therefore, there is a definite requirement to investigate the methods and places where data mining in E-learning can be employed, exploited and used efficiently in addressing the peculiarities of E-learning processes.

This thesis examines how data mining can enhance the quality of the learning process. The case study of this thesis is to use data mining to predict students' performance and to assess its effect on the overall achievements of the student. Students' performance prediction is an important device for enhancing the learning process in E-learning systems (Gagnes 1997; Wang and Shao 2004; Chen, Chen, and Liu 2007).

It is well recognized that in face-to-face learning, instructors build experience in predicting their students' performance and, hence, can exploit ways to improve and optimize their instruction accordingly. This is not similarly possible for E-learning due to the remote communication and individual nature of the process. Therefore, this thesis investigates whether data mining can overcome such difficulties and substitute human experience in predicting student performance.

Many mutual benefits exist in the use of digital technologies for learning, both for the instructor and the student (Aldhafeeri & Male, 2015; Courville, 2011; Liaw, 2008).

Instructors can devote themselves to teaching, understanding the factors influencing students' performance, tuning teaching strategies for students with various learning performance factors, early identification of students at risk of underperforming, enabling the instructor to put in place remedial action well in advance and also assisting the instructor in identifying in advance students likely to qualify for scholarships and fellowships. Online feedback mechanisms incorporated into the E-learning system can greatly assist students in understanding their current status and make adjustments accordingly if required (Krause et al., 2009). This also helps with motivation and a feeling of general inclusiveness as opposed to isolation and the tendency to 'drop-out' (Krause et al., 2009).

Conversely, E-learning systems can inadvertently reduce the need for interaction between instructors and students whereas, in traditional classroom systems, instructors can monitor the progress of each student and can, accordingly, predict how the student's achievement would be affected by his/her current progress. This is not practically achievable in E-learning environments, and this is something that needs to be addressed.

In this thesis, it is hypothesised that using data mining techniques on the archived Learning Content Management Systems (LCMS) data can help in assessing the current student progress and, hence, predict his/her level of achievements based on experience gained from historically archived similar progress data of those colleagues who have passed the same course in previous intakes. Accordingly, students can be automatically advised by various strategies of interaction with the LCMS with the goal of achieving better results. E-learning students are unlike traditional students in that they have many unique circumstances affecting how they can improve their learning performance. The main objective of this thesis is to develop a system from the instructors' point of view as a virtual mentor and build a prediction model that predicts students' performance based on their usage of the system. The results produced from the model will be used to enhance the E-learning system by sending a message to each student describing his/her performance; these messages would improve the students' performance and encourage them to do better.

SPS is a prototype that is built for this thesis using Java. This programming language was selected by the researcher because there are many open source machine learning and data mining tools, which have been implemented in this language which should help to achieve the research objectives.

1.2 Research Background

Technological and scientific progress has led to this area being characterised by change. Consequently, it is also necessary to ensure that this pace is the same to that of educational processes through facing challenges that may happen or arise. These changes may include an increase in the number of learners, information increase, distance and lack of teachers, etc.

Numerous studies have been conducted to determine the extent to which E-learning is beneficial. Many researchers came up with strong evidence, which showed that E-learning can result in reduced learning time, and also reduces cost as training within this area is extremely easy compared to instructor-led training (Raspopovic et al., 2016).

Since its emergence, E-learning has encountered several problems such as the management of large amounts of unstructured information; if such situations are left unsupervised, they could easily lead to system deterioration. As a result, to prevent such problems, data mining techniques should be used. Data mining refers to a collection of methods for data analysis, where individuals who want to evaluate information make use of an automated or semi-automated process that constitutes pre-processing, data understanding and modelling.

This procedure is used to help in bridging various traditional fields which are different (such as pattern recognition, statistics, etc) to help in providing the individuals with a variety of analytical solutions towards solving specific problems. Educational institutions in all areas are also provided with the ability to focus on the most important information in databases and also explore it. Using exploration techniques will help to focus on determining future predictions and explore trends and behaviours that allow the making of timely correct decisions.

1.3 Research Problem & Questions

There is a need for a more adaptive E-learning approach to give more accurate prediction and better advice for the students. This kind of adaptive approach is supposed to go deeper in comprehending a student's style of learning and the student's history as well. The goal of this research is to come up with knowledge that improves systems of E-learning predictions in higher education. The first stage of the research was to help explore the ever wide range that exists in E-learning systems which use data mining in predicting the performance of students.

Knowledge is always growing, and consequently, the interest of the research continues to focus on how to develop a new system for the prediction that will greatly benefit and utilise data mining to predict students' performance and advise them on how they can best improve their marks. This has led to the research problem of this study: "which mining techniques are the most suitable for this domain that can be applied to help develop an effective advising system that enables students to improve their performance?"

The research will contribute to improving the recommendation system by enhancing its abilities and allowing it to better serve the needs of the users. The main emphasis is laid on the integration of the different approaches so that the overall performance can be improved. The addition of the features that are related to the learning styles of the learners should be beneficial. The learners will be having an advantage of focusing on the appropriate activities and achieving better performance.

As a result, the questions to be answered by this thesis are:

- Is data mining usable in the environment of E-learning?
- Can the E-learning environment learning process be enhanced by data mining?
- Can the performance of students be predicted?
- Can this prediction be used to help students improve their performance by using the E-learning system?
- Can this prediction be used to help the instructor enhance his teaching process and provide appropriate advice to students?

1.4 Research Objectives

The main objective of this research is:

> *The application of data mining to data available in the E-learning environment so that the E-learning process can be improved by predicting students' performance and suggesting advice to both the instructors and students based on the predictions made.*

Accordingly, the approach and roadmap followed by the research were as described below:

- Consider the necessary background on the data mining algorithms, data mining techniques and tasks.
- Construct necessary background on all the data mining tools used.
- Explore and analyse popular E-learning systems.
- Choose the data mining method that is most suitable.
- Design a system for student performance prediction (SPS).
- Assess the SPS prototype through the use of anonymised Dammam university student samples.
- Provide recommendations for students and predictions for instructors by using the SPS prototype.

1.5 Definitions

A number of terms were used frequently throughout this thesis, and for clarity's sake, definitions of those terms are provided here.

- Data mining – "Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database. The discipline concerned with this task has become known as data mining." (Hand, Mannila, & Smyth, 2001).
- E-learning – "eLearning is about using information and communication technologies (or ICT) to expand access to education and to enhance and transform teaching and learning." (Bullen, 2014).
- Performance prediction – the process that allows one to "foresee", "forecast" or "predict" the performance of a certain individual, based on already collected data. Data mining is carried out on the data and used for performance prediction. Here, to define "student performance prediction" is simply related to the current thesis.

Student performance is given by grades, and the prediction of these grades is done by the proposed system. The prediction is the proposed system's output that forecasts the expected results of each student.

- Advising system – a computer-based system (software) that has the ability to produce advice based on collected data about the individuals (in this case, students). This means, generally, taking multiple factors into account – again the input is obtained through data mining of student data – and coming up with an automated response that is useful for the student. Generally, academic advising takes place in "situations in which an institutional representative gives insight or direction to a college student about an academic, social, or personal matter. The nature of this direction might be to inform, suggest, counsel, discipline, coach, mentor, or even teach". (Kuhn, 2008, p. 3).

1.6 Scope of Research

Data mining and E-learning existing knowledge are used as the basis on which to conduct the research. In chapter 2 there are reviews of the areas of data mining and E-learning which are the main focus areas. This research's main scope is the development of E-learning software that incorporates data mining techniques. Nevertheless, there are wider implications to the study, particularly:

- Improving E-learning through a deeper understanding of the correlations between student individuality, work undertaken and results achieved
- Improving the accuracy of prediction software by making use of data mining
- Gaining insight into the effects of personalised student advice
- Further understanding and improving virtual learning environments

Acknowledging limitations such as a limited data set and population or limitations in sampling, the scope of the study revolves around further understanding student performance in a Virtual Learning Environment (VLE), studying the impact of advice on student performance, and improving on previous prediction models.

1.7 Research Methodology

This research is purposefully done to propose the best or most effective way to enable students to predict their performance and also to advise the students on they can improve it to achieve results which are better. Hence, the research is done out in the positivism and objective paradigm, fitting in perspective of experimental research. The methodology for the research is divided into four phases: analysis and data collection phase, planning phase, system development phase and the conclusion and discussion phase. The table below illustrates these four phases.

Table 1 Research process illustrated

Planning Phase
Define Objectives, Research Problem and Scope (Chapter 1) Research Areas and Review of Literature (Chapter 2)
Phase of Data Analysis and Collection
Research Methodology (Chapter 3) Design and development (Chapter 4)
System Development Phase
Evaluation & results (Chapter 5)
Discussion & Conclusion Phase
Discussion of: Research Output, Limitations, Contribution to Knowledge, Future Work (Chapter 6)

1.7.1 Planning Phase

This phase studies the topic of research and identifies gaps in the literature. The objective, research problem and the scope were defined initially. For better understanding of the topic, the main areas of research – data mining and E-learning context - have been reviewed. According to the literature review, it has been revealed that several studies in data mining and E-learning have been conducted before. These reviews are found in chapter two. Unknown situations and gaps have been identified at the end of the review.

1.7.2 Data Analysis and Collection Phase

In this second phase, the research method is designed based on the research problem and literature gaps found from the planning phase. A pilot study was also conducted, then the researcher moved to the main phase of the study. This study has indicated that the best way to collect data is by using the existing knowledge that can be found from multiple sources. This research is conducted with an experimental research strategy using multiple methods of data collection. This is chosen as the pilot study has indicated that data are difficult to collect due to their private and confidential nature.

1.7.3 System Development Phase

Most of the activities are carried out in this phase. Testing and validation of the software are also done at this stage using a sample of Dammam university students. The number of students is one hundred. Reports are also produced which are similar to what the students will get if they were using the real system.

1.7.4 Discussion and Conclusion Phase

This is the last phase and reflects on the developments. The discussion here addresses how the study has achieved its research objectives. The conclusion will summarise the study's main findings as well as analyse the research contribution. This phase thus aims to distinguish between findings in order to establish the best ways of improving student performance. To do so, it will analyse and interpret results, discussing their implications and the measure in which they answer research questions. Finally, the phase has the role of improving knowledge in the field by building on previous research and providing additional data into the application of data mining in E-learning processes.

1.8 Thesis Organisation

The research is categorised into six main chapters, and they are as described below:

Chapter 1 covers the introduction, which will describe the problem definition of the research, motivation and the research approach of how to tackle the problem. This chapter aims to introduce the reader to the context of the research as well as develop a preliminary understanding of the research problem, the objectives of the study and the methodology designed as to meet these objectives. Additionally, it describes the phases of the research and the structure of this report.

Chapter 2 involves the review of the research work where data mining is used for E-learning. Additionally, it presents the required background involving the data mining process, concepts, techniques and models. This chapter also details the context of this research, enabling this study to identify gaps and formulate research objectives. As such, the literature review is critical not just in establishing the academic background, but also in placing the study in a wider context of research, to be used and improved on in further research.

Chapter 3 is about the research approach and methodology which are used by the researcher in conducting the research. The chapter discusses the philosophical underpinning of the research, its characteristics and the methods it uses to achieve its goals and answer research questions. Moreover, this chapter discusses issues of reliability and validity, as well as research limitations, enabling the reader to understand both advantages and limits to the study.

Chapter 4 considers how a decision tree learner (ID3) can assist the student's understanding. This chapter discusses previous work in this area, the approach used in the research, the advantages and disadvantages of the prediction systems and lastly evaluation of the existing systems. In addition, Hidden Markov Models (HMM) are used to help recognize the students' learning style. This chapter also discusses the use of an ontology to build models of student knowledge.

Chapter 5 describes the implementation and design of prototypes developed for the evaluation and testing of the model. The conclusion and evaluation resulting from the systems development are also included. This chapter is important to better understand the process and results of the research, describing the performance of the prototype and its implementation.

Chapter 6 is the final chapter, that presents conclusions as well as implications in enabling further research. The chapter also aims to further discuss data mining in E-learning, describing how this study improves knowledge in the field and how E-learning can be improved.

Chapter 2 Literature Review

2.1 Summary

In this chapter, the main topics, and relevant areas will be introduced which have relevance to this work. First, an overview of data mining will be given including its main methods and common techniques. After this, E-learning will be discussed, including its advantages and disadvantages and how progress has been made from instructor-led training through to adaptive E-learning. Finally, the researcher will introduce adaptive learning, its main concepts, and how one can consider learning styles in such systems.

2.2 Data Mining

Data mining was developed in order to tackle the problem of discovering patterns in large data sets. Defining data mining is difficult. It is not because of its fundamentally complex nature, but because it has most of its origins in the ever-shifting world of business. Partially, it is a set of methods used for data analysis. It is furthermore a data analysis process that incorporates anything from understanding data and preprocessing and modelling it to evaluating and implementing processes. (Castro, Vellido, Nebot, & Mugica, 2007)

Data mining methods bridge the fields of traditional statistics, pattern recognition and machine learning (ML) to present analytical solutions to queries in areas as different as biomedicine, engineering, and business, and many additional domains. These techniques can be used in E-learning through the analysis of the information available about each student to understand their learning behaviour pattern and the suitable learning paths for them.

In this section, two of these techniques will be discussed in more detail.

2.2.1 Preprocessing the Data

In this section the focus is on preprocessing the data, and on why preprocessing would be necessary. As the literature reveals, certain efficiency requirements need data to appear in certain ways. More accurate algorithms often depend on the quality of the methods used during the preprocessing stage.

The first step in preprocessing is making sure the data is in digital format. For this purpose, human entry and scanning are used. While scanning with OCR (optical character recognition) is a powerful tool, it requires human entry as well. It is ideally known that all data should be consistent, accurate and complete, yet people face challenges with real-life situations. Databases usually do not have the characteristics of how information should be stored, thus data mining is used.

To use data adequately, one must make sure that the data carries the necessary quality. Inaccurate data, outdated data, and not easily interpretable data would all be wrong choices. It is essential to understand what are or what could be the causes for lower quality data.

To begin with, larger databases storing an almost uncountable number of values are prone to more mistakes. A simple incorrect human entry (e.g. writing ‘dear’ instead of ‘deer’ or other typos), a runtime error of the software that was used to save the changes to the database, stored items categorised incorrectly (e.g. adding ‘milk’ to ‘IT Department’) and more can cause low quality. As manual review is not possible for large databases, errors need to be tackled strategically. It is a good idea to learn what the best ways are to deal with errors in large datasets. Ignoring the tuples that have the errors is rarely a solution, while completely eliminating errors may be highly challenging.

Normalising data can also become important at certain stages. Assume there is a database of different sizes of certain parts, all being measured in cm, m, in, feet, and various measurement units. In that context, normalisation means that all measurement units are converted into a standard unit of choice. This process provides more accuracy.

Certainly, the purpose defines what data quality is. If a particular data set is enough for a certain task, the same data would make some other task impossible. One can easily take name, email and birth date to keep track of people and acknowledge their age groups, such as under 20 years, 20-30 and more, and would allow easy email notifications sending. Having only name, email and birth date on the other hand will not be sufficient in the case of tracking the same people’s shopping habits from a given online store.

Frequently people deal with untimely entries, or not having the necessary entries on time. The combination of all the mentioned elements calls for Data Cleaning, a process that is very important.

2.2.2 Data Cleaning

2.2.2.1 Introduction

Once the challenges in preprocessing data are understood, the importance of data cleaning is undoubted. In this section, the data cleaning methods are discussed as well as real-life situations when cleaning is necessary and approached differently.

2.2.2.2 Cleaning

Encountering noise in data tuples from any database is unavoidable. The problem is that the noise in the input may cause problems in the output, especially if the learning process would use parts of the input where some essential information is missing. When there is input containing noise that needs to be considered, the primary stage is detecting the legitimate noise that is part of it (García, Luengo, & Herrera, 2015). Cleaning only makes sense after noise is identified, and the post-cleaning task is constructing the algorithm that will implement the learning process that would not be affected by potential remainder noise in the data.

Joining multiple entries based on the level of similarity has been presented by (Chaudhuri et al., 2006). While there are methods that can be used (by comparing strings) to determine how similar two entries are to each other, there is no final "best choice". The research in the publication further proposes the introduction of SSJoin, which is similarity joins of entries based on certain sets they have attached to each entry. SSJoin analyses the similarities between two given sets. To understand better, one can consider strings. These strings are broken down into 3-grams (for example the word "understanding" has the following 3-grams: und, nde, der, ers, rst, sta, tan, and, ndi, din, ing) and then for the two different strings, SSJoin will first count the number of matching 3-grams. Having the numbers, one can have full overlap (at least 80% of the smaller length), 1-sided overlap (considering 80% of the first string) or 2-sided overlap (considering 80% of both strings) (Chaudhuri et al., 2006). Edit distance, Jaccard similarity and generalized edit similarity stand at the base of the efficiency of SSJoin. Edit distance measures the operations required to repeatedly edit a given string, turning it into another given string. From 'cats' to 'cat' the edit distance is just one, because the letter 's' will be erased. Jaccard similarity would break down strings into multiple groups of three characters, but will not accept matching of two groups unless the groups are the same. To enhance the functionality – or in other words the lack of optimal matching – of the two above mentioned functions, the generalized edit similarity associates a cost to specific string operations, thus better assessing the similarity between strings.

All RFID (Radio Frequency Identification) systems will rely upon the available data cleaning quality. RFID allows identification by radio frequency, meaning this is possible from a given distance. Want (2006) presents the efficiency of RFID tags, but also argues on the expenses to replace barcodes (as people have seen in stores, on any product) with RFID tags. The main advantage is the possibility of identifying large amounts of data from a distance. Gonzalez et al., (2007) take on the problem of data cleaning for RFID while considering costs. The presented approach is designed to minimise the costs, thus resulting in a more efficient and practical system. The algorithm is based on Dynamic Bayesian Networks which are similar to Hidden Markov Models, (Murphy 2002). The Dynamic Bayesian Networks use multiple random variables for both the hidden state and the observation. All of these random variables are either continuous or discrete. Additionally, these scholars have measured several types of cleaning costs, and used their own cleaning plan for testing the proposed solution.

Dallachiesa et al. (2013) presented NADEEF. NADEEF is a platform that can be used to clean data, and it is both general and easy to configure. The mentioned environment is divided into a core and a programming interface. With the programming interface, it is possible to define rules for acceptable data quality. Then the core will deal with the defined rules, applying the necessary algorithmic approaches. The researcjer tested the proposed environment to prove that it is effective, general and also extensible.

Measurable Data

Measurable data is generally all data stored as floating point numbers or integers. The complexity of measurable data can be anywhere from simple numbers to complex multidimensional number arrays with multiple parameters. Regardless of the complexity, the cleaning technique that is broadly used relies on statistics. In other words, this method checks the entries that are valid. Based on those entries, it is easy to determine which values seem to be unusual or improbable. This method is called Outlier Detection. (Hellerstein, 2008)

Postal Address Data

Hellerstein (2008) also discussed data cleaning approaches on postal addresses. The article presented the two most significant issues in this area. The first issue has to do with the way a certain address is written (e.g. “23 Main” or “Main St, 23” or “23 Main Street”), and the second difficulty is the spelling of names (e.g. “Steven Jackson” or “Stephen Jackson”). There are commercial solutions to clean postal address data.

Manual editing

While manual editing is possible, it is not recommended due to inefficiency. Manual input would take more time and effort based on the size of the database (Han et al., 2012)

Probability or Mean Values

Han et al., (2012) also discusses the possibility of using probability or mean values computed from stored values. In other words, if the values are already known where not missing, there are algorithms to determine the probable value. To establish correct mean values, it is important to first understand the classification of the data as available. For example, it is incorrect to use the mean price of dairy products to establish a missing price on a dining table.

2.2.3 Data Reduction

2.2.3.1 Introduction

In this section, the role of data reduction is discussed. The section considers why it is important and how to reduce large data sets efficiently. The researcher also goes through parts of the literature, each focusing on the different aspects of data reduction.

2.2.3.2 Data Reduction

Today, databases can grow to enormous sizes for many applications, thus analysing entire databases for certain purposes becomes nearly impossible. For this reason, it is important to reduce the original data set into a smaller set that yields acceptable results for the various goals.

The main question is: would one approach be able to apply data reduction ideally, without any loss? In rare scenarios, for certain purposes, it is possible to have an ideal reduction. For example, if there is a large database containing student data from all high schools in the U.S., it can contain dozens of fields per entry. At the same time, if the purpose was to find out whether education quality (based on grades) has improved at a certain high school, the reduction is already enormous. In other instances, ideal reduction is not possible. To be more precise, most real-life situations would demand a certain, minimal data loss.

The goal of quality data reduction is assuring the minimal loss, but keeping the results close enough to what the original complete database would have yielded. Feature selection is an efficient data reduction approach, as discussed by (Dash, 1997). Starting from an input feature set – that is to be reduced – a potential subset is generated. Choosing the right subset can be done by various methods, such as randomly selecting features or simply including all features and gradually eliminating features. The generation, once complete, needs evaluation to assess whether the potential reduced data set can become the current best reduced set or not. New subsets are generated to ensure that the most reduction is taking place. A well-defined stopping condition must be included to stop the process of generating new subsets. Once the cycle is completed, a validation process assesses if the resulting reduction is valid. While all of the above steps have their specific algorithms and procedures, it is not the scope of this thesis to discuss all the details.

One notable work in this domain is related to the Karhunen-Loeve Expansion and how it can be applied in feature selection (Fukunaga, 1970). Phoon and Quek (2002) define the Karhunen-Loeve Expansion as follows. “Basically, K–L expansion provides a second-moment characterization of a random process in terms of deterministic orthogonal functions and uncorrelated random variables… The deterministic eigenfunctions are obtained from the spectral decomposition of the covariance function.” Further research was conducted on this topic by multiple scholars, such as (Jensen & Shen, 2007). When selecting certain subsets, one must make sure that the original attributes are effectively described by the attribute subset. While it is possible to analyse all subsets in an attempt to find the best choice, it is not efficient – it uses a lot of memory and takes too much time on even small databases, and its application becomes impossible for very large data sets. Thus the best approach is a greedy algorithm, which captures the first of what looks the most promising, using a tree-like approach in which it assesses potential best subsets. In other words, a greedy approach will not check if its choice was indeed the best, but it will assume it is good enough.

Another known technique is reducing data numerosity. The methods within this reduction technique will replace the original data by some alternative data that can represent the original. Histograms and sampling are just two of the well-known techniques. (Han et al., 2012)

The most widely-known data reduction technique is called compression. Here, an alternative representation of the original data is used that doesn't require as much storage. If the representation can be reverted to fully reconstruct the original, it is considered a lossless compression. A lossy compression means that the original data will not be recovered entirely, but only estimated (Han et al., 2012).

Other scholars, including Wickerhauser (Wickerhauser, 1994) presented the principal components analysis method (PCA). PCA is used in data reduction to reduce the dimensionality by transforming the original data into a lower-dimensional representation. It is good to acknowledge the fact that PCA will likely, in most cases, reveal hidden aspects about the data. This results from the construction and methodology of the algorithm, the actual step-by-step PCA. However, the original meaning of the data can be lost due to PCA being a transformative approach.

Histograms are also known for their broad usage, and are discussed in many articles. In 2016, there are notable works that relate to histogram usage and data mining, such as (Wang, & Li, 2016; Quellec, et al., 2016). The first of the mentioned works deals with spatial data mining, in which it is discussed that histograms can stand at the base of further knowledge, if the used samples in building a feature histogram are sufficient. Li et al. also presented strategies to use histograms in the mining of temperature data. Quellec et al. uses histograms for mining visual words, as a study on retinal pathology. Then the scholars conclude that using one histogram will not solve the general problem, thus a combination of methods was accepted as a working solution.

Essentially, a histogram approach is a representation of counts per occurrence. For example, in a city, a histogram would reveal how many people are between the ages of 20 and 30, how many are between 30 and 40 and how many are above 40. Thus, using such a reduction allows one to know the number of people within a certain age group without counting the people's names, one by one (consider extra time spent when going through a database of a few million people). While one may often think about one dimensional histograms that concern just one parameter, the literature also employs multidimensional histograms, as in (Bruno, 2001). The scholars discuss how based on the dependency of more attributes, a multidimensional histogram is highly efficient. The discussion demonstrates that it is unnecessary to process the data fully, as the multidimensional histogram is based on results of queries.

In this thesis, data cleaning plays an important role. While multidimensional histograms are not required, feature selection and reducing data numerosity are important parts of the presented work.

2.2.4 Clustering

Clustering is a way of grouping together data samples that are similar. It is a form of unsupervised learning and data exploration for patterns and structures of our interest. The concept of distance is considered as an essential component for clustering techniques. Regarding the known categorizations of clustering methods, the following four exist: Hierarchal clustering, Partition clustering, Grid-based clustering, and Density-based clustering.

One can also observe clustering in papers that present student modelling topics. One of the notable articles is (Amershi and Conati, 2009). The scholars describe a comprehensive picture of the presented classification. They also define what clustering path was chosen to benefit user modelling. In a learning environment, they have tried to give more exploratory than traditional tutoring systems, as students are required to have a deeper and a more structured understanding of the concepts in the domain. Exploratory tutoring systems rely on the ability of the students to explore the subject matter, rather than having a fixed structure of what, where and how to begin. Traditional tutoring, however, is better for some students but it does not limit the potential of the freedom of exploration as part of studying.

Amershi, et al., (2006) describe another similar approach in which clustering is used to be aware of learner groups in the exploratory learning environment in an automatic way. The approaches adopted by the authors are focused on identifying the student behaviour for online learning by using unsupervised and supervised classification of students and works with logged interface and eye tracking data. This is an entirely different approach from the ones the researcher is trying for classifying the students and providing the content based on real-time performance.

The clustering approach which is based on the collaboration behaviour is given by Anaya and Boticario, (Anaya & Boticario, 2009) describing how the statistical indicators in learning activity data are used for determining the cluster membership.

2.2.5 Sampling

Sampling as a data reduction technique is meant to choose an adequate sample to represent the full data set. This is often done by randomly picking a defined number of tuples that are going to form the reduced data set (Han and Pei, 2012). While choosing other approaches may require more memory and time, sampling is fairly simple and fast. The focus must be on how large a sample should be in order to stay within a defined error margin (error allowance). If one considered a database containing millions of records, a sampling of one thousand records could be efficient for certain tasks. If accuracy must be fine-tuned, sampling size can later be adjusted according to the task's demands.

At first, the development of sampling was not for computer use, but rather for experimentation. However, for the current thesis, it is only relevant to present the modern data sampling approaches which have emerged along with contemporary technology.

Monte Carlo technique

Presented, among others, in the 2009 book of Lemieux (Lemieux, 2009), the Monte Carlo technique is not merely about data sampling. This powerful technique, applicable to a diversity of different fields, is used today as a reliable means of addressing various problems.

Monte Carlo, first of all, is based on arbitrary inspection. With the help of this tool, one can analyse the characteristics of a setup, given that units have arbitrary actions. The technique works with the help of computers, where the setup actions are simulated. During this simulation, arbitrary action descriptors are generated. Based on the output of the aforementioned, it is possible to statistically interpret the results of the analysis.

Later studies, as will be reviewed in the following subsections, brought about the desire to have something more powerful than the arbitrary inspection and interpretation given by Monte Carlo approaches.

Deciding the Scope of the Model

In data reduction, as part of data sampling, an important decision is the scope of a selected model. If a model includes the required scope, but also goes beyond it, then the data reduction can be deemed incomplete. Also, if a model is incapable by scope to represent all necessary aspects, then the data reduction was exaggerated and will not be useful to determine or measure characteristics based on the partial input given. Thus, a 2002 study by Muthén and Muthén, (2002) revealed how Monte Carlo can assist in such decisions, enabling one to choose the right scope for a given model.

When working with Monte Carlo, it is crucial to assess the tendency of standard errors, but also of criterion approximation. Inclusion is equally important in assessment. The following decisions then should be made:

- Using multiple seeds (in the context of randomness)
- How many patterns to use
- Selecting the prototype

The scholars opted to examine two different models. One of them was CFA and the other was Growth, which will be briefly described in the next paragraphs.

CFA (Confirmatory Factor Analysis) as a model has been used by scholars in various tasks over the last decade. Based on the brief description presented in Schreiber et al., (2006), an estimated model is required which means two matrices are used: one is noticed and checked (measured), and the other is approximated. Both matrices are covariance matrices. The goal is gaining a minimal variance or, in other words, maximal similarity between the two. Thus, there is an association between what is measured and what is not measured, but only estimated – this association is observed as part of the CFA model, as described above.

Secondly, Muthén and Muthén, (2002) made a use of two growth models, one having a covariate and the other not having it. As discussed in Durlauf et al., (2001), the growth model referred to in the 2002 research is known as the Solow Growth Model. The model can provide stable interpretations, which includes development proportions. While the original concept of the prototype was meant for economics, contemporary studies have taken the Solow Growth Model to other domains.

2.2.6 Probabilistic Models

One of the best known probabilistic models is the Hidden Markov Model (HMM). The first mentions of the HMMs are in: (Stratonovich, 1960; Baum and Petrie, 1966; Baum, et al., 1970). The HMM is considered a highly efficient mathematical means, used to model productive arrangements. With these models there is a hidden mechanism that generates noticeable series. Many works have used HMM such as Soller & Lesgold (2007) and Soller, (2007). Soller and Lesgold described the modelling process for the example case of knowledge sharing, defining the knowledge sharing episode. They define it as a multitude of talkative inputs and activities. Such knowledge sharing episodes begin as soon as anyone from the organization brings contemporary awareness into the discussions of the organization. The episode ends as soon as there is no more talk regarding the contemporary awareness. Again their usage is quite different from the way the researcher is using the HMM models in the described system. For efficient content delivery, it is required to provide the content to an individual student as per his/her learning style preference. As discussed in the previous section, students have a bias towards one of the three learning styles, but students tend to jump from one learning style to another from time to time. To model efficient and intelligent content delivery model, a Hidden Markov Model approach is used in the present work.

Another pattern detection approach is described by Beal et al., (2006) who used HMMs to model students' performance on problem-solving. Later on, the models become a basis for prediction in the process. The HMMs becomes a well-fitted analytic approach for being used for explicitly modeling and indicated by the better prophecy accuracy, compared with the simple Markov chains. Their approach is more focused on assessing the engagement levels of students for a mathematical tutoring system, and in general not used to provide content as per the performance and engagement levels of students.

Jeong and Biswas, (2008), presented an approach to behaviour modelling. A study is described with middle school students operating with Teachable Agent (TA). The proposed TA is a software system created to improve the way students learn. It must be trained, which is achieved via the student through a GUI (Graphical User Interface). Then, the current knowledge of the TA must be improved. This process of improvement is done by supervising the current knowledge of the TA and correcting it whenever it gives a wrong answer and the student notices this. They also use HMMs to represent the sequences of activities in order to make known the patterns which lead to learning success.

Li and Yoo, (2006) explained the modelling of student learning performance with Bayesian Markov chains which were used in the adaptive tutoring system. Referring to the work Li and Yoo presented, but also to what has been done in (Sebastiani, Ramoni, Cohen, Warwick, & Davis, 1999, pp. 1-2), the “Bayesian Markov Chains” is about a Bayesian algorithm that is capable of learning Markov Chain representations, finally clustering data as the task demands. Sebastiani et al. used this algorithm for modelling time series by categorising them into clusters based on time series that show similar dynamics. Li and Yoo presume three basic student models which are based on the following three learning types: the reinforcement type, the challenge type and the regular type. According to the author the challenge type is the student who learns from his mistakes, the reinforcement type is the student who needs repeated reinforcement to be able to understand a concept and finally the regular who has both the previous two types of pattern.

2.3 E-learning

Since the first public school in the US, Boston Latin School, was founded in 1635, technology has dramatically emerged over the years to become one of the main aspects in today's world of education.

From pencils and school books to the internet and modern technologies, it is easy to clearly note the effect of new technological inventions in learning styles and methods; for this reason, many researchers divide the history of education depending on the inventions which lead to new "educational revolution." For instance, Billings and Moursund, (1988) cited four revolutions in the history of education. These revolutions are the invention of reading and writing, the emergence of the profession of teacher/scholar, the development of moveable type (print technology), and finally the development of electronic technology which leads to E-learning.

The following sections will focus on the "fourth educational revolution" and how E-learning becomes one of the important axes in the learning processes.

2.3.1 Open learning, Distance Learning and Online Learning

Learning by courses sent via regular mail was first heard of in 1840, invented by Sir Issac Pitman. Prior to 1969, distance learning was already advanced in many different countries. In 1969, in England, a novel approach was proposed that combined distance learning with media and phone-call tutoring. This took place at the British Open University. These concepts were introduced not to eliminate traditional universities, but to enhance them, giving countless people the opportunity to benefit from higher education (Matthews, 1999).

One of the first notable works on open learning comes in the year 1973 (Wedemeyer, 1973), when the researcher focused on the open learning system and also on what conditions would allow open learning to even occur. While there have been several mentions of open learning mostly as "open education" before 1973, these were either lacking solid foundations or focused on principles that upon a closer inspection prove to have nothing in common with open learning (for example: an educational method focusing on the tutor).

In the year 2002 UNESCO introduced a platform for open learning. Thus, they have defined open learning as “Open Educational Resources” as a public access to educational resources that can be changed and used freely, but not sold. (D’Antoni, 2009, p. 1). One can see the emerging of open learning through the above definition (specifically targeted to be part of “Open Education Resources”), or simply the idea of people to make certain study materials available on the Internet, or in libraries. Good examples of open learning sources could be Wikipedia and Ehow, two websites widely used by people of all ages.

Distance learning requires no physical attendance, but rather the enrolment into a university or other institution online, receiving open learning materials through the Internet and taking exams also online. The concepts of distance and open learning are intertwined; open learning makes less sense without distance learning, and distance learning is difficult without open learning. There has also been research that considers how students should be encouraged in these areas (Simpson, 2013).

Online learning is a general term and can refer to any activity to learn any type of material online (i.e., by using the Internet). These concepts of open learning, distance learning and online learning are all part of E-learning, which is presented in the next section and defined in detail. Online learning has been part of academic circles, but people also learn online on their own without being part of any academic program.

2.3.2 Relationship between instructors and students

As presented by Frymier and Houser, (2000), the relationship between the instructor and students is similar to that of friendship, with a few differences. The mentioned relationship is discussed by scholars, because it directly affects the quality of learning.

According to Muller (2001), there is a certain importance associated with instructors that are caring for their students versus the ones who either do not care, or their students do not feel like they cared. As shown by Muller, students who feel judged or told that they cannot succeed will most probably fail or pass but with minimal effort or interest. Thus, it is easy to understand that while instructors are the ones who show the learning materials to the students, the relationship between the instructors and students will also impact performance.

2.3.3 Concepts of andragogy and heutagogy

Andragogy has been well-described as a concept in Knowles (1970). Briefly, andragogy is the process of helping adults with their learning process. While the approach of what andragogy is in completion of mere pedagogy (pedagogy as a concept was also described by Knowles and many other scholars), it is important to see beyond these older concepts.

Andragogy brings the following advantages:

- A straightforward approach to offer education to individuals (not necessarily taking into account how personal learning experiences would be different for the participants), based on the ability of the teachers to prepare necessary materials and teach the relevant subjects in a structured way.
- In a teacher-centered environment, andragogy is very natural and simple for the teacher.

Andragogy, however, has disadvantages as well:

- It does not consider learner styles, learner personality and is overall not learner-centered.
- It does not let the learner experiment or "learn how to learn", but instead enforces its own logic, structure and approach. While some subjects such as mathematics have one correct solution for many problems, there are different approaches that lead to the same solution, which is the same with learning. One can learn something in many different ways.
- It does not consider the information era, meaning the multitude of channels through which information is accessible and thus learning is possible.

Heutagogy shares a new point of view – that is the importance of the student in the process of learning, meaning that the student should find a proper way to learn, rather than being directed or forced to learn in a certain way. So, in other words, heutagogy is assessing self-determined learning (Hase and Kenyon, 2000).

Advantages of heutagogy:

- Considers the importance of the student
- Encourages self-determined learning
- Enhances learning quality

Disadvantages of heutagogy:

- If certain students would be highly disorganised or requiring special attention, heutagogy would not enable them to learn anything.
- A lazy person could also end up not learning with the heutagogical approach.

2.3.4 Constructivism, connectivism, or both?

While there are differences between constructivism and connectivism, it is good to know about the basic principles of both approaches.

As presented in Siemens (2014), constructivism is building up knowledge by producing meaning, instead of merely having no knowledge and waiting for knowledge to be poured in from the outside. While this principle addresses the on-going, experimental process of learning, it is limited as it does not take organisational knowledge and transference into account. Siemens, in the same article, stated that connectivism addresses the constantly changing nature of information beyond any control of the learner in many cases. Thus, a connectivist learner will not rely on the same foundations every day, but rather focus on any connection that allows further learning, knowing that the new information can change everything compared to yesterday.

It is fair to say that for now, the principle of connectivism is up-to-date, taking our society and technology into account. Each connectivist learner should also be constructivist – in the information expansion all around, one should always produce meaning based on new information, and only secondly on what was available previously (if it does not contradict the new information).

2.3.5 Impacts of digital technology on the learning environments within institutions

Digital technology definitely changes environments, as many fields are digitised and virtualised. In relation with the current thesis, it is important to assess the impacts of digital technology on learning environments within institutions.

Collaborative learning is hard to define, but Bruffee (1999) gave a more general definition. One can say that collaborative learning means a group of people who are gathering knowledge together. In the non-digital sense, one would assume "together" means meeting face-to-face. Thanks to digital technology, "together" can mean any of the following: Skype conferences, emails, Face Time sessions, WhatsApp or similar applications (chatting), Facebook messaging, SMS, phone calls and much more. Learning, in the classical sense, could mostly mean academic learning or learning from life experiences. Digital technology allows learning from or with: software applications, wiki sites, online forums, digital books (or any other digital material available for studying) and more.

Then, self-directed learning is something that relates to collaborative learning at some level. A simple definition is given by Brookfield (2009) as follows: “Self-directed learning is learning in which the conceptualization, design, conduct and evaluation of a learning project are directed by the learner. This does not mean that self-directed learning is highly individualized learning always conducted in isolation. Learners can work in self-directed ways while engaged in group-learning settings, provided that this is a choice they have made believing it to be conducive to their learning efforts.”

Digital technology changes self-directed learning, given the fact that information is now available digitally in abundance, to the extent that a person could not assimilate it all. Thus, ultimately, the used sources and resources will define the quality of the acquired information, and digital technology also facilitates multiple means of communication. This also results in going through these learning processes faster, due to the digitisation. However, it also means more effort in obtaining more information in the same amount of time (compared to how information gathering was before the digital era).

2.3.6 Defining E-learning

It is known that the term “E-learning” stands for describing learning through electronic means, or shorter electronic learning. It further describes how technology can be included in education, as a valuable tool. However, since the first use of the concept “E-Learning”, scholars introduced various rationale. Nichols described E-learning related with high-tech means. He stated that these means should be either distributed on the web, based on the web or able to integrate with the web (Nichols, 2003). Other researchers have proposed more general definitions to cover content, and instructional methods delivered not only via the Internet, Intranet, or CD-ROM (Benson et al., 2002; Clark, 2002) but also via audiotapes, videotapes, satellite broadcast and interactive TV (Ellis, 2004). One of these general definitions has been proposed by Gilbert and Jones, (2001) who described E-learning as a distribution of learning materials. They established the distribution as through any electronic means including audio/video tape, the Internet, satellite broadcast, interactive TV, CD-ROM, computer-based training (CBT), and intranet/extranet.

In fact, there are many other definitions, but most of them –especially the general ones- focus on a common quote used to describe the final goal of E-learning. The quote is “Anytime, Anyplace, Any pace learning” (Birch and Clements, 2003), where “Anytime” relaxes the need of the teacher and the student available in one common time. Meanwhile “Anyplace” means that people do not need to gather in one single place. Moreover, In “Any pace learning”, different students can follow the course at their own individual pace. A fast learner can finish the course in a shorter time than a slow learner.

Some other terms are also used to describe this E-learning technology. They include online learning, virtual learning, distributed learning, computer-based learning and web-based learning. In general, they all refer to educational processes that utilize information and communication technology to improve learning and teaching activities. These terms could not be deemed as equivalents for E-learning because most of them cover only a part of E-learning directions (Alghamdi, and Jensen ,2014). Let us consider what distance learning is. The preliminary condition that allows distance learning is that the involved students do not attend any class physically. Thus, deciding by what means content should be delivered is crucial for success in E-learning practices. Still, this is just one aspect of E-learning.

2.3.7 E-learning Advantages and Disadvantages

E-learning has both advantages and disadvantages. It is not suitable for all the conditions but offers many flexible and creative benefits for teaching as well as learning. There is a dire need for realising the advantages and the disadvantages as a safe way of dealing with any technology in order to avoid problems and any arising difficulties.

The Advantages

E-learning encourages the development of independent learning environment. Furthermore, E-learning brings fifth benefits for tutors, but also for students. Here are the advantages that could be observed (Itmazi and Megías, 2005; Najafabadi and Mirdamadi, 2013):

- *It is flexible and convenient*: It is accessible anytime and anywhere. Also, users can select the learning materials depending on their interests and needs. They can access various resources, do quizzes and tests and get immediate feedback, and they repeat lectures and videos many times easily. All these points reduce the stress and increase the confidence and satisfaction. For teachers, E-learning is easier for monitoring the progress which can be more precise and less administrative work. It enhances the quality of teaching and can be assessed systematically and evaluated electronically.

- *It saves time and cost*: E-learning provides flexibility in terms of time, space and speed. It provides a more efficient learning experience. If a person is busy with their other commitments, such as family or work, then E-learning helps in enhancing education in parallel. An individual who has any type of disability and is not able to attend the classes can also take the opportunity to learn through E-learning. It is cost effective because the user can repeat the course without spending extra money or fees. In the case of any difficulties learners can participate in many discussion forums or threads and they can also keep track of student’s knowledge area so that they can find expertise of the student.
- *It is a learner-centred environment*: The learner is the only beginning and ending point throughout the learning. In the process of studying, the learners are the operative participants. Their requirements stand in the centre of attention.
- *It is consistent*: It provides the same quality of content and education for everyone. Moreover, it is useful for students of different learning and cognitive styles. On the other hand, E-learning provides a solution for the teacher shortage problem, as it is known in many disciplines that this is a problem.
- *It is interactive*: E-learning improves the interactivity between students and teachers, especially in large lectures and it promotes collaborative learning.

The Disadvantages

Studies proved that such tools are not suitable for every situation, so there are some challenges in E-learning, such as Itmazi and Megías, (2005):

- *Lack of face to face interaction with teachers*: Students may feel isolated. Meanwhile, the discussion forum reduces this problem, but it is sometimes not possible to meet face to face due to the disarray of students' physical locality (Pivec, et al., 2004). Also, the present human-computer interaction is still lacking the emotional component, thus providing the learner with the feeling of being treated in an impersonal way (Pivec and Baumann, 2004).
- *Technology dependent*: Applying E-learning needs strong technology infrastructures including computers, fast internet connection, and reliable network. These

requirements increase the start-up cost and make it difficult to apply E-learning in some countries.

- *Unsuitable for some cases*: some courses rely heavily on interpersonal contact. While these courses could be complemented by E-learning, they are not brought forth well enough by using E-learning. First, any learner must have a high amount of motivation to press through. Only with enough motivation are there advantages in using E-learning for studies. Second, the motivation does not guarantee success. Discipline and managing time wisely are both essential if a student desires to achieve success with any given E-learning course.
- Relying on assessment marks may involve some imprecision and uncertainty.

2.3.8 From Instructor-Led Training to Adaptive E-learning

Instructor-Led Training (ILT) (Kapp et al., 2009) was the primary training method when computers were not widely available. The technological advancement of multimedia attempted to make learning more transportable and visually engaging. The "anytime" and "anywhere" availability of CD-ROM provided time and cost saving, but it lacked instructor's interaction.

As the web evolved, the appearance of email, HTML, Web browsers, media players, simplistic Java and low accuracy streamed audio/video began to transform the face of multimedia-based education. Technological advances including Java/IP network applications, rich streaming media, high-bandwidth access, and advanced Web site design are revolutionising the learning industry. These sophisticated learning solutions provide even greater cost savings, higher quality learning experiences and are setting the trends for the educational standards of the future.

The early E-learning systems suffered from fundamental problems such as the lack of a substantial degree of personalisation that can occur at the content level. In particular, systems do not take account of prior knowledge, nor do they account for any differences in the learning style (Coffield et al. 2004).

With the rapid development of Web technology today, E-learning evolves to become adaptive E-learning where the focus is more towards the individual learning environment and the application of self-regulated education. Many techniques and methods have been designed to support learning in more adaptive environments. Every method presented above brings forth both weaknesses and advantages (Kinshuk and Patel 2003). The aforementioned methods handle the main issue of adaptive E-learning. The main issue is the problem of user assessment. Due to the lack of accuracy, the issue is further the way to display the system's user consciousness and continue to update it. In reality, distributing users between categories of which each describes a hobby is not the issue.

Up to this moment, most of the current platforms providing distance learning through the internet have a knowledge content which is not being adapted to the particular needs and styles of the individual learners; generally E-learning platforms provide content which is uploaded prior to the course being offered online. Therefore, the content been provided may or may not contain performance evaluation from time to time, and do not have closed feedback loops to adjust the content delivery models as per the performance scores of the students. To evolve into adaptive learning, researchers are tackling these issues using an adaptive algorithm for content delivery taking into account the individual learning styles of students and providing a real-time feedback loop and subsequent adjustment of content delivery models based on the performance of students from time to time (Downs et al., 2001; Arellano et al., 2004).

2.3.8.1 Adaptive E-learning

The adaptive system is a subfield in machine learning which relies on the information they get from interaction with the user to provide more accurate and useful services next time. By processing the collected data, the adaptive system would be able to adjust its own configuration and operations in acknowledgment to the feedback from its environment. Furthermore, it would draw some conclusions of the user's needs and requirements (Ahmed et al. 2005).

In general, the classical machine learning techniques have been used for collecting, interpreting and processing of the user activity and interaction data. The semi-intelligent techniques are necessary to get hidden information which could not be extracted using simple statistical techniques.

Adaptive systems are used in many applications such as E-learning systems, financial forecasting, fraud detection, credit ratings, robotics, plant modelling and control, digital cameras, medical diagnosis, search engines, and computer games.

2.3.8.2 Definition of Adaptive E-learning System

To define what an adaptive E-learning system is, it is best to understand that one of the major criteria is having a system that can adapt to personal needs. In other words, if a student has a certain personality, an adaptive E-learning system would easily bring forward the available study resources by taking the student's personality into account. Today, the adaptive learning system is one of the most important issues in E-learning; an increasing number of conferences and workshops focus on this topic. Researchers (Hawk & Shah, 2007; Neuhauser, 2010) show that when students learn through their preferred learning strategy, they will take less time to master a given concept in a more comprehensive way.

There are further details regarding the adaptive E-learning system. One can separate the kinds of adaptation as discussed by Martinez (2009): First there is adaptively and second there is adaptability. The distinction between these is only whether the system takes action first or the student. Concerning adaptively, the system takes action. In this case, the system identifies certain traits of the user, and based on these assumptions it will comply in an automatic manner. Adaptability means the student takes action first. This also means that the users could modify given parameters through the system, making sure that the user behaviour is adapted. When a definition of characteristics in an adaptive E-learning system is given, it must suggest a balancing of the mentioned kinds of adaptation.

2.3.8.3 Challenges of the Adaptive E-learning System

There are many challenges to designing adaptive E-learning systems (Shute et al., 2003). The most important ones are the following:

- The learning content and learning concept should be suitable for each individual's strengths and weaknesses.
- The organisation of pre-tests to judge the knowledge background of learners.
- Intelligent Tutoring Systems (ITS) are very complex. They are also time consuming and expensive.

- The course authors, the tutors and the developers of the E-learning environment must be aware of the pedagogical principles.

2.3.9 Adaptive Learning and Learner Styles

In this section, learner styles and models are discussed, and then proceeds to relating the learner styles to adaptive E-learning systems.

2.3.9.1 Learner Styles

Learning styles affect the education methods, concerning the individual that allow the individual to learn better. One starting point is the belief according to which everyone has preferences related to information or even stimuli. The preferences are at the level of intake, processing and also interaction. According to the above stated claims, in the 1970s researchers proposed personalised learner styles. Today the same principles started gaining more interest. One of the core recommendations or strategies targets the methods used by teachers. It is better for teachers to understand the students' learning styles first. Once the styles have been recognised, the methodology and technique of the teacher should change and mold into what is best for the student, based on their studying habits.

The learning styles usually benefit from many research results of the cognitive psychology. Vainionpää in 2006 describes the styles. The scholar stated that the majority of the students will favour certain representations of information. Furthermore, their other preference is a well-defined response over other students, as soon as they have a distinction in learning. Due to how every student has a multitude of different styles of learning, they blend these styles in an attempt to get the best possible combination for every episode of studying.

One of the most significant aspects is how students can recognise their own learning style and any effect it has on the learning process. As Coffield et al. (2004) say, knowledge of learning style can be employed to increase students' self-awareness and the metacognition of their strengths and weaknesses as learners. Merrill (2002) says that the majority of students are noticed to be unaware of their learning styles. When such students are left without guidance with their style questionnaires, they may be not able to start learning in new ways. Coffield declares that such students who lack confidence in their learning ability can be motivated to look for new ways of exploring and describing their behaviour as learners.

A 2002 study (Neuhauser, 2002) analysed whether one can gain the same efficiency by teaching online and by older methods, where learners would attend physically and listen to their tutors in person. To ensure efficiency in determining results, learners were divided into two different groups. The groups were given the same study materials along with the homework assignments. The online group was never physically with any of the tutors – they used e-mails and E-learning. The offline group did meet physically and had nothing to do with E-learning; the contact was only through e-mails. Factors such as past experience regarding the use of technology or online platforms in studying, as well as age, years of employment (if any), and gender were analysed. While slight differences were observed between the offline and online groups, according to the conducted chi-square assessment the study revealed no major discrepancy. The author thus measured the preferred learning style of each student and how efficient the learning is with the chosen styles. The study also has a significance for E-learning, because it also measured teaching through the Internet. In reference to Learning Style Models (see section 2.3.11.3 – Learning Style Models), the two mainly encountered favourite learning styles were the kinesthetic and the visual in both groups. The conclusion of analysing the association between achieved grades and style preferences was that there is no association. Furthermore, there was no difference between the online and offline groups. The only small aspect of distinction was that the online team achieved somewhat better results.

Positive aspects of distinguishing learning styles

In 2011, as presented in the work of Gilakjani (Gilakjani, 2011), there are positive aspects when one can distinguish learner styles. It is noteworthy how one’s own correct assessment of learning preferences can aid the individual in speed and quality of studies. Dissatisfaction mostly results from studying subjects that are not related to the student’s learning behaviour. On the contrary to this, a correct assessment propagates success. Self-control while learning is also necessary, but impossible without complete knowledge of one’s learner style. As a conclusion, by distinguishing the styles, students learn how they need to study. This promotes the feeling of accountability and consciousness, as in being aware of their own actions, preferred styles, potential study results. Lastly, without properly assessing styles, students tend towards unknown goals of their lives and careers. If their learning styles were properly identified, the students will have clear goals – which are self-assessed once the style is pinpointed – and overall better achievements in life.

2.3.9.2 Relationship between Learning Styles and Teacher Styles

Today scholars are interested in the relationship between learning styles and teacher styles. The main focus is on whether the students can benefit from certain teacher styles and if that is possible, it is important to determine how it will happen.

A study by Gilakjani in 2012 compared these two broad subjects. While for each teacher it is vital to have extensive knowledge of their subject matter, it is equally important to inspire confidence and carefulness as well. Furthermore, a professional tutor will not only possess wisdom in their own area, but also intelligence of using proper teacher styles and insight to notice each student's personality. (Gilakjani, 2012)

It has been stated that as soon as any tutor starts teaching, they come with a philosophy of their own, whether they know it or not. In conclusion to different possible approaches that teachers may or may not use, the effectiveness would be in making sure that the class has more interest, engagement and overall quality of studies. Not every class responds to the same styles, and not every subject matter allows the same tutoring methods. The scholar found that a suggestion in this area is targeted towards the teachers. Tutors should be able to play with a palette of styles, thus assessing which style is in best relationship with the personalities and preferences of their students.

The researcher concludes with various important statements regarding the relationship between student learning styles and teacher styles. One description is that the first important step would be correctly assessing one's own learning style. As mentioned in the paragraphs above in this section, the self-assessment will bring various benefits. Then, on the other hand, tutors should also be able to evaluate what their psychology is, and also what style they use. If tutors can mould their styles to the personalities and demands of the students, the overall study experience should improve and result in academic enhancement regarding grades, students' personal satisfaction, but also tutor's professional achievements and satisfaction. Often it is only possible by teamwork to assess the tutor styles and when each of them would be appropriate for implementation in class.

While studying the EFL (English as a Foreign Language) and ESL (English as a Second Language) classrooms (Iranian students), the researcher proposed a number of promising practices for tutors, to use in class:

1. Engage students by talking, asking and answering questions

2. Provoke attendants by contemporary queries (ask novel, challenging questions)
3. Introduce ingenuity by inventive exercises
4. Discuss prototypes that will prompt inquiry and integration
5. Assess both student styles and own styles, engage in talks about classroom proceedings
6. Provide a diversity of assessments to strengthen talent and variety
7. Provoke students to practice views, theories and make use of available knowledge
8. Use a blend of audible, ocular, physical and palpable methods
9. Adjust methods and find tasks that relate to students' styles

There is a novel study on the subject, conducted in 2015 (Rogowsky, et al., 2015). The scholars assessed the learning preferences of students participating, and they established whether learners are ocular or auditory. While they state that many researchers believe that best achievements come only when the tutoring style is adapted to the student's style, they made a contrary statement based on the conducted experiments. While comparing the two groups of students, they observed that the ocular group would achieve more on auditory tests. As a conclusion, they suggested further research and an open-minded attitude towards the possibility that some student styles should rather be reshaped into other, more efficient styles (e.g. auditory students should be assisted to gain more visual skills).

Obviously, to accept the 2015 findings there needs to be far more research and discussion on the topic. For example, this research is brief and only selects a group of individuals with certain studies finished and a certain age. As found in the paper, the average age of the participants was around 30 years old. Thus, the first problem arises: how can one correlate the findings on mature adults, who have completed their studies, with the findings on how students (young people still studying) would respond and learn? Several other influential factors from the chosen age group have been omitted, such as:

1. Work experience gained – which is also learning.
2. Preferred learning style – it was not clear whether these were the remembered preferences of these people (such as how they considered to have a preference maybe 10 years ago) or these were current preferences (as in the boss requests them to learn something new, and they go about it in a preferred learning style)
3. How would students respond to this same type of study, what results would it yield?

Thus, for now, these 2015 findings lack depth and also lack the ability to definitely and unquestionably assess learning styles in relation to students.

2.3.9.3 Learning Style Models

Various learning style models are available in coherence with different aspects and requirements. There are many ways of looking at a learning style. In this section, some of the most significant models are described and discussed. Some of the noticeable models are the following:

- VAK Model
- Kolb's Learning Model
- The Study Standard presented by Mumford and Honey
- Prototype proposed by Felder and Silverman
- Model proposed by Dunn and Dunn

Visual-Auditory-Kinesthetic (VAK) model

The VAK model focuses on the pathways of human perception. These pathways are feelings, vision and hearing. In the research of Sarasin in 1999 (Sarasin, 1999) it is clearly concluded that students can be categorized as Auditory (A), Visual (V) or Kinesthetic (K), thus the VAK abbreviation. The classification is completed based on the individual choices of processing information. Thanks to this simple approach, the VAK model still has scholar interest until today, being widely used in academic research. The visual perception or personality comes from a favour towards anything that presents knowledge through diagrams, images, tables and many more. Other students will rather hear about their lessons and get involved in any discussion, thus being auditory in perception. Lastly, the kinesthetic students are the ones who always enjoy experiencing information, by touch, movement and exploring. The only issue is classifying the students correctly based on their learning styles. A methodology commonly used for classification will include questions which purposefully target daily behaviour, response in given situations, and generally questions helpful to determine the students' personalities.

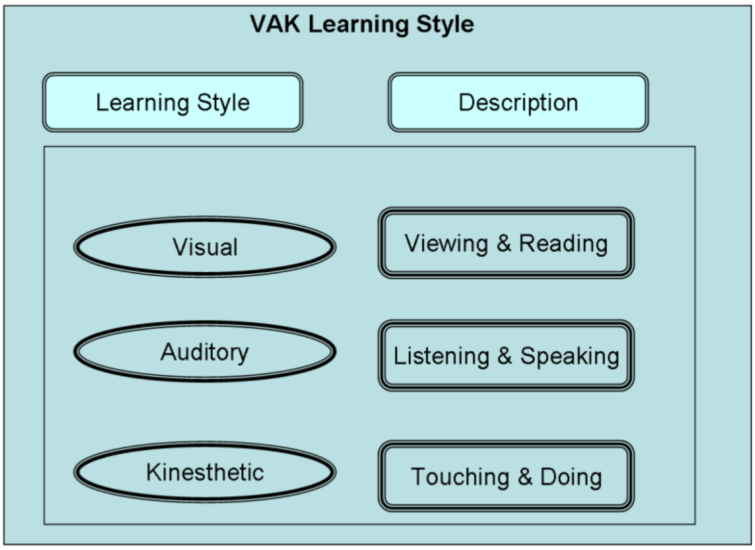


Figure 1 VAK Learning Style Mode

Scholars in 2008 (Sharp, et al., 2008) debated the importance of the VAK model, and discussed many details related to multiple learning style models. The research states that it is over-promoted as a learning style. Furthermore, while learning is a non-trivial process that is not fully known to researchers – there are still aspects and distinguishing ways learning happens within each human being – and the VAK model would suggest that it is far simpler than reality. The scholars also claimed that it is merely a redefinition of what people already know about how children and students generally learn. As a model, the following weaknesses are assessed and documented: sufficient diagnostic data is not available, prediction is minimal or non-existent, and there is almost no didactic force in this model. The lack of didactic force comes from the fact that the model only presents some essential patterns of how children understand things surrounding them (including everything they learn). Since this is something known without reference to the VAK model, the model alone fails to bring any didactic force. Further on, since the model only makes a theory of how children respond to teaching and interaction, the idea of prediction is impossible. The model is not a prediction, it's rather a statement of facts. Lastly, one does not yet have diagnostic data to use in relation with the VAK model.

Kolb's Learning Model

Another well-known model is Kolb's learning model. This developed style model was published in 1984. Four learning styles were originally included in Kolb's learning theory. All these styles are based on a learning cycle of four stages as shown below:

1. Concrete Experience (CE)-Feeling

2. Reflective Observation (RO)-Watching
3. Abstract Conceptualization (AC)-Thinking
4. Active Experimentation (AE)-Doing

Learners absorb these reflections and then translate them into abstract concepts such process enables to create a new experience and also helps to start a new cycle.

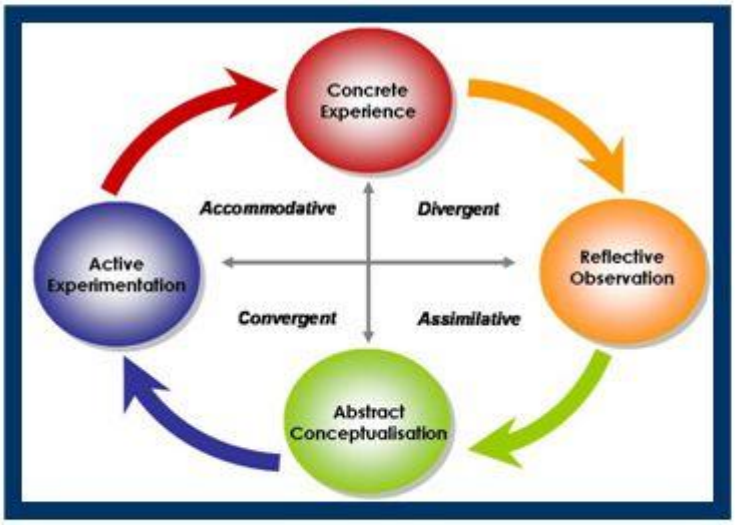


Figure 2 Kolb's Learning Style Model presented by Atkinson, A.G. (2011)

As shown in the above figure, the learning cycle in this process is represented in the areas of learning and where the experience, reflecting, acting and thinking are treated. Definitions of the learning style are not more than mere representation of combination of the two styles that gain learner's preference:

- Diverging (CE/RO): A process of learning through emotion and watching is known as diverging learning style. Students that have a diverging personality in learning will rather enjoy visual matter than action itself. These learners are usually better at viewing situations from several different viewpoints. They prefer group work. According to the research found in (Smith, 2001), these students will enjoy collecting knowledge. They also observe individual criticism, and have a tendency to rely on inspiration when issues need to be solved.
- Assimilating (AC/RO): Watching and thinking as ways of learning, are combined through the assimilation learning style. The interpretation of the word assimilation is grasping the interpretation (English-Finnish General Dictionary 1997). They need clear explanations and prefer logically sounding theories over those who are based on

practical value. This kind of learners prefer readings, lectures and exploring analytical models (Smith, 2001).

- Converging (AC/AE): Learning by doing and thinking is the converging learning style. The convergers are seen as they like to combine ideas and practice closer together. This category is not concerned with interpersonal issues or people because they prefer technical tasks. They prefer experimenting new ideas to imitate and practice applications (Smith, 2001).
- Accommodating (CE/AE): Doing and feeling is the way to learn for the accommodating learning style. The word accommodation means adaptability here.

The mentioned student learning style takes insight in favor of rationale. The learners who follow this style take a practical and experiential approach towards the learned material. For completing their tasks, they prefer to work in teams (Smith, 2001).

Additionally, this model is strong and has been modified, adapted and highly useful over the years. There are mentions of using Kolb's Model in education. It is also a reliable model, and researchers confirm that in multiple findings, such as (Kiili, 2005; Corbett, 2005). The contemporary disadvantages linked to this subject are: no real conclusions in scholar claims, contradictions related to proving it is accurate. (Sharp, et al., 2008)

The Study Standard Presented by Mumford and Honey

A contemporary standard for study styles was presented by Mumford and Honey, in 1982. It was based on the work of Kolb though it was different, and it included the four key stages:

- Activist
- Reflector
- Theorist
- Pragmatist

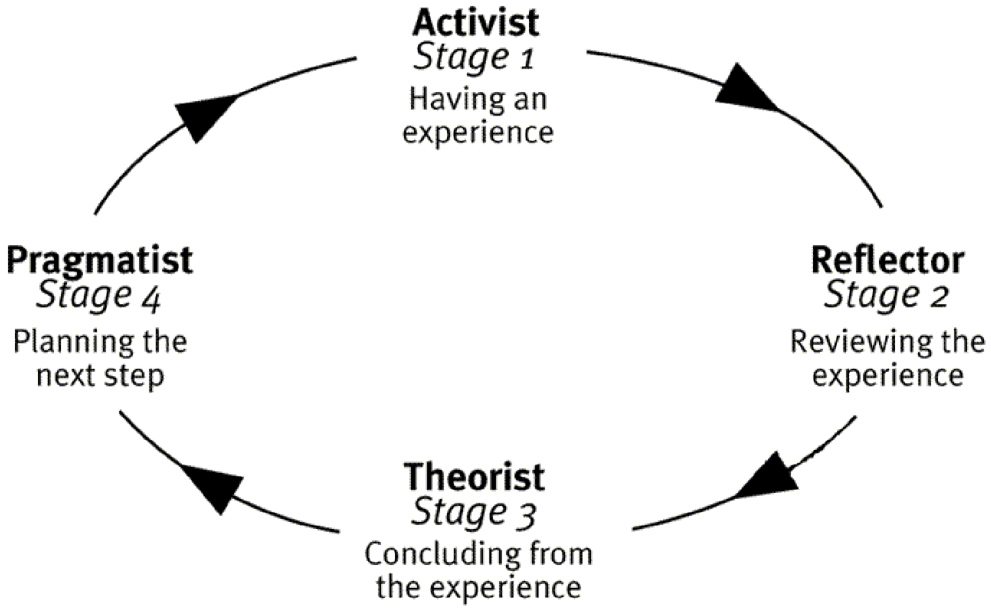


Figure 3 Learning Style Model presented by Honey and Mumford's.(1992)

The above four key stages represented mainly Kolb’s styles of learning though they have slight differences. Their preferred way of learning is group activity; they work with tasks and some educational games. When they listen to lectures, read and write on their own, these activities obstruct the activists' learning. They do not like to follow the precise instruction or the strict schedules (Campaign for Learning, 2006).

The style stages of Honey and Mumford are almost adding together with the Kolb's learning styles. Here is the correspondence between them:

- Activist will correspond with Accommodating
- Reflector represents the Diverging style
- Theorist is the same as Assimilating
- Pragmatist means the Converging style

This prototype they developed is also known as “Learning Styles Questionnaire”. Scholars debate topics such as trustworthiness and legitimacy in relation with this model. (Sharp, et al., 2008).

Prototype proposed by Felder and Silverman

This prototype can be associated with the names of two scholars: Richard Felder along with Linda Silverman. In 1988, these researchers proposed the student’s learner style are referring to as the “Felder-Silverman model”. This is created for engineering students. The prototype bears the name of these two scholars, and is also known as the FSLSM, which means the "Felder-Silverman learning style model". It has four elements and classifies students as:

- Anticipative or Perceptive
- Optical or Lexical
- Operative or Thoughtful
- Continuous or All-around

The anticipative and perceptive students desire to study evidence. They also get solutions for queries by using recognised procedures. Meanwhile, the intuitive tend to the discovering possibilities. The active learners prefer trying things or doing something active. Additionally, thoughtful students focus on having their own mindset and thoughts related to any matter. Continuous students will prefer studying by going forward one step at a time. All-around students also comprehend subjects in steps, but they may easily go more steps at a time (Felder, 2002).

According to (Carver et al. 1999), the FSLSM is the most pertinent technique for didactic setups placed in computers. Learners are asked to express their personal preferences for each dimension with values between (+11 to -11) per dimension with steps (+/-2). All this is done via a 44-item questionnaire. This range comes from the eleven posted questions for each dimension (Carmo et al. 2006; Graf 2007).

Kuljis and Liu (2005) compared several learning style theories and assessed whether or not the studied models are suitable for E-learning. In their research, it was found that the Felder-Silverman model is the most appropriate for E-learning. Others have referred to the Honey and Mumford model in their study to create a unified learning style model (Popescu et al., 2007). The second research, targeting the development of the ULSM (Unified Learning Style Model) took the best of all available learning style models, but only the characteristics that would:

1) Prove a major influence throughout the studying process
2) Could be adapted to the needs of an online learning system
3) Are assessable from student observable behaviour in an online learning system

The research came up with the so-called WELSA (Web-based Educational system with Learning Style Adaptation) that implemented the ULSM principles. (Popescu, 2010).

Dunn and Dunn – a famous work on a well-known model

In academic circles regarding the available and widely-spread learning style models, the work of Dunn and Dunn is one of the most well-known. Based on the research of Hawk and Shah (2007), the Dunn and Dunn approach mentions five catalysts regarding study behaviours:

1. Indirect
2. Sentimental
3. Social
4. Cognitive processing
5. Physical

Based on the catalysts, there is a questionnaire to be answered which will determine the learning style based on the responses. Each of the five catalyst groups brings several aspects to it, allowing a more precise measurement. Scoring for the Dunn and Dunn model is very straightforward: between two options there are three score groups: 20-40, 40-60 and 60-80. The 20-40 group means the student preference goes for the option at the left-hand side; 40-60 means there is a uniform preference between the presented choices; 60-80 indicates favouring the right-hand side option from the questionnaire.

Sharp, Bowker and Byrne (Sharp, et al., 2008) also compared the strengths of this model with its own weaknesses. This model is approachable for users, having personalised options for people of different ages and presenting subsections as well. It is also a combined model with academic application mainly in the United States. The authors Dunn and Dunn (1978) emphasised how important it is to match learning style with the student’s learning preferences. Throughout the literature this has been debated, with some scholars saying that the learning preference indeed facilitates better learning, while other scholars stated that the learning preference has little to do with results and performance of students. There is no separate assessment to confirm or contradict the usability and correctness of this model. Critics stated that the model also contains oversimplified statements along with lack of complexity. According to criticism, such as (Felder, 2005), there is not enough applicability of this model.

2.3.9.4 Learner Styles in Adaptive E-learning Systems

As the main purpose of the adaptive system is to provide more suitable learning environments for the learners, it is important to deal with an appropriate learning style model. For this reason, many researchers try to integrate different learning styles into adaptive E-learning systems.

Surjono approached the application of the Adaptive E-learning system in (Surjono 2009). He proposed using both the VAK and Felder-Silverman prototypes, integrated in Moodle. Moodle is a widely-used system that facilitates student data management, track records (courses, grades, attendance) and is freely available as an open-source solution, easily adaptable to the requirements of any tutors in terms of style, layout, etc. In 2011, further research came by Sharif and Mustafa (2011). Their paper analysed the possibility of a newer path towards the integration of learner styles in an adaptive E-learning information bank. They also discussed about certain outcomes resulting from personalizing the study materials based on each learner's style of perceiving information.

As early as 1991, Riding and Cheema classified building learner style as either analytical and wholist or imager and verbalizer (Riding and Cheema, 1991). The category that has the analytical and wholist learners best analyses how students handle information. The analyst style means that these are people who enjoy seeking minutiae within information. Wholists on the other hand will study anything from an overall point of view. The group with imagers and verbalizers analyses the expression of knowledge. Imagers are students who display any knowledge in visual forms. At the same time verbalizers will use words to describe what they know or want to say (McLoughlin, 1999). An older mention of the group with wholists and analytical students is found in Pask's research (Pask, 1988) where he describes serialists and holists. Pask stated that serialists are students who can easily follow detailed instructions. Wholists would first want to review the materials and only after the review would follow into detailed steps. Felder and Silverman stated something similar, but described the concepts from another point of view (Felder and Silverman, 1988). They said wholists are global, serialists are sequential, imagers are visual and verbalizers are verbal. The sequential student group would follow continuous steps and then reasonable bit-by-bit patterns. Opposed to sequential, global students would rather study while skipping continuous path bits.

Most of this research shows that the students who were taught by using the adaptive learning style system performed significantly better in academic achievement than those who were taught by the same material without adaptation to the learning style.

2.4 Conclusion

Reviewing the literature is eye-opening about the existing methods, learning styles, data mining concepts, E-learning advantages and disadvantages. There are still multiple areas in which things are going to change, as well as areas that are always under debate with some scholars supporting the respective area and other scholars trying to prove it wrong at the same time. Thus debates are questionable, until certain scholars will choose to prove certain facts right or wrong by solid evidence and research.

The literature review also allowed choices to be made in terms of used methods, system design and overall improvement to the existing literature, in parts strictly related to student performance prediction in E-learning.

The following chapters will present the algorithms and models of choice, the construction of the proposed system and the measurements and research conducted to prove the benefits of this system. The final aim is to help both teachers and students via the currently proposed solution.

Chapter 3 Research Design and Methodology

3.1 Introduction

In order to achieve an algorithm that can predict student performance with accuracy, there is need for clear, reliable and valid research design. This chapter aims to outline the challenges of the research, the sampling process, how research was carried out and how results were collected and analysed. In the process of achieving this, it will also discuss the philosophical research implications, the ethical implications and the timeline of the research as well as the theoretical background underpinning the research design.

The research follows a quantitative approach through experimental research in an attempt to answer research questions with regard to the capability of predicting student performance. The research is carried out in two steps: developing the Students Prediction System (SPS) algorithm based on statistical analysis of historical student performance, and testing the algorithm's predictions on 100 students during a semester at the University of Dammam in Saudi Arabia.

The reasoning that underpins the design choice of experimental research based on a quantitative design is that these methods best allow the researcher to investigate research questions and meet research objectives. As a result, this chapter analyses the theoretical underpinnings of both, together with their practical implications. For instance, Wiersma & Jurs (2005) argue that experimental research must include four critical elements: manipulation, control, random assignment as well as random selection. This chapter will discuss the practical implementation of each, while also discussing associated research validity and research limitations.

3.2 Research Objectives and Questions

The study's aim is to apply data mining techniques as described in Chapter 2 to the E-learning environment at University of Dammam in order to predict student performance, offer students advice based on the predictions, and then analyse the impact of the advice on student performance. The impact of the advice will be analysed in comparison to a control group which will not receive the advice.

First, the study identified several research objectives and defined a number of research questions. The aim of the objectives is to clearly outline the aspects that need to be investigated and enable the researcher to develop a roadmap to achieve these goals.
The identified research objectives are:

- Building the necessary background regarding data mining tasks, techniques, and algorithms.
- Building the necessary background regarding data mining tools.
- Studying popular E-learning systems.
- Selecting the most suitable data mining method.
- Designing the (SPS).
- Evaluating the SPS prototype by testing it on a sample of students.
- Applying the SPS prototype to provide predictions and recommendations for students and instructors.

Underpinning the research objectives are the following research questions:

- Can data mining be used in an E-learning environment?
- Can data mining enhance learning processes in E-learning environments?
- Is it possible to predict students' performance?
- Is it possible to use this prediction in helping students use the E-learning system better and improve their performances?
- Is it possible to use this prediction in helping the instructor provide appropriate advice and enhance the teaching process?

Additionally, two research hypotheses will be formulated:

- ➢ An SPS model based on data mining can predict student performance results in a relatively accurate manner, underpinned by statistical research.
- ➢ The advisory component of the model can help students achieve higher performances, by allowing them to benefit from personalised advice based on the prediction model

These hypotheses will be directly tested by this research study.

3.3 Research Strategy

This chapter will outline the research strategy, describing its development, the theoretical process underpinning the experimentation, as well as the steps and procedure of conducting this. First of all, this research will employ a quantitative, experimental research strategy. Making a choice regarding research strategy must occur naturally, as a result of the research objectives, as the research strategy must represent the most effective method of achieving these (Creswell, 2013). As the primary objective of the research is developing an effective SPS prototype, it becomes apparent that the best method of verifying results is by testing the prototype on real data, through an experiment. Since the advising component of the model is critical, because of its ability to interact with students and enable positive change, it was decided to test the prototype in a live scenario by attempting to predict future performance, rather than testing it on historical data. While testing the model on historical data would have been simpler and less time consuming, this choice of research design has the advantages of allowing the researcher to gain insight into the practical applications of the prototype and into the effects of providing personalised advice to students regarding their performance. Therefore, there is a dual role for the experiment:

1. Observing the effects of personalised advice based on statistical data to students.
2. Testing the actual performance prediction capabilities of the algorithm, in particular testing whether the algorithm can achieve significantly better accuracy than a basic neutral stance (such as assuming that the results will stay the same over time).

To further justify the choice of research design, the study will briefly look over inadequacies, advantages and disadvantages of other research methods. Firstly, it can be noted that qualitative research is inadequate in meeting the research objectives, since the research requires an objective and measurable research result. The only aspect where qualitative research could have been useful is in better understanding the subjective experience of research participants in interacting with the advisory software. However, this was beyond the scope of this study, even though it is an area that shows potential, especially as data mining and E-learning continue to develop. Secondly, other research strategies such as questionnaires or archival research simply do not meet all research objectives.

With regard to the actual experimental procedure, this research project consists of several main steps:

- Gaining access to university data, which was achieved through correspondence, and was particularly time-costly.
- Researching data mining techniques and deciding on a model based on the Hidden Markov Model.
- Designing and developing a software prototype that will predict student performance in an E-learning environment, based on variables such as the amount of time spent or the number of quizzes solved. The development is made through the use of historical data collected from students at the University of Dammam.
- Testing the prototype during the course of one semester by providing advice resulting from performance prediction to a group of students, while the other students of the same course serve as a control group.
- Collecting final performance data, analysing it, and discussing results and conclusions.

3.4 Research Philosophy and Approach

An important aspect of the research lies in the philosophical assumptions that underpin choices in the research design and methodology. Research philosophy is concerned with the way in which the researcher perceives the nature of the world and the nature of knowledge (Holden and Lynch, 2004). These aspects eventually lead to the formation of systems of logic, which are critical to how the study attempts to meet research objectives, depending on how the outside world and the data being researched is perceived (Gray, 2013). This chapter will outline the philosophical basis of this research and attempt to establish a logical flow that takes research design decisions from ontology to epistemology, approach and methodology.

Ontology is the philosophical branch that deals with the nature of existence and reality. Ontology therefore asks questions regarding fundamental meaning, identity, properties and, most importantly, how these can or cannot be identified (Holden and Lynch, 2004). The ontological perspective of this paper is objectivism, which operates under the assumption that the nature of the world is the same for all entities, and does not fundamentally differ depending on the observer. This has critical implications and enables quantitative research by not considering results as the subject of interpretation, but rather as subject to understanding through scientific methods.

Epistemology on the other hand deals with the nature of knowledge and the constituency of facts (Gray, 2013). Essential to epistemology is what true knowledge represents and the methods through which this knowledge can be achieved. For instance, Descartes made the famous affirmation “cogito ergo sum”, which represents the only aspect of reality of which humans have epistemological certainty: that thinking is the only objective proof available for our own existence, as all other evidence is gathered through subjective means (Holden and Lynch, 2004). This also highlights the unbreakable links between ontology and epistemology. The epistemological assumption adopted by this research is a positivist rather than interpretivist one, in accordance to its objectivist ontology as there is a natural pairing of ontology and epistemology: objectivism leads to positivism, while constructivism leads to interpretivism. Positivism is the epistemological belief that positive knowledge, truth are objective, clear, measurable and can be verified through scientific, reliable methods (Gray, 2013). Positivism is therefore based on natural law, logic and reason.

Furthermore, looking into the research approach, epistemology allows to discuss the way in which theories are developed and truths are established. This approach can either be deductive or inductive. Deductive approaches develop hypotheses based on theory, and then attempt to observe and confirm the validity of said theories (Gray, 2013). Inductive research on the other hand follows observations in order to reach hypotheses and develop theory. This research follows a deductive research approach: the primary assumption of the study is that data mining can be harnessed to improve student performance. Based on this approach, hypotheses were also constructed: the SPS model’s ability to accurately predict performance, and the advisory component’s ability to help students achieve higher performances. These are direct results of the deductive method, a defining feature of quantitative research. An inductive research approach would have instead studied the general phenomenon of E-learning in order to reach conclusions and formulate hypotheses about the field.

Table 2 Research Considerations Summary

Research consideration	**Paradigm and Type**
Ontology assumption	Objectivism
Epistemology assumption	Positivism
Strategy	Experimental
Methods (Data collection and analysis)	Literature Software design

3.5 Sampling

One of the most important aspects of quantitative research, of particular importance in experiments, is the sampling process. Sampling is essential in ensuring that the participants of a study are representative of a wider population, and therefore this particularly impacts the reliability and validity of a study (Cohen et al., 2013). Sampling is therefore the method through which large groups are researched in studies where interacting with the entire population is impossible or inefficient (Cohen et al., 2013).

The effectiveness of the sampling process is primarily determined by two aspects: "the size of the sample as compared to the size of the population, and the sampling method of choice" (Saunders and Lewis, 2012, NP). While there are clear mathematical formulas determining confidence level and confidence interval, a more difficult choice regards sampling methods. An important distinction between sampling methods is made by their probabilistic vs non-probabilistic nature (Saunders and Lewis, 2010). Probabilistic sampling methods are those that utilise random selection methods and include stratified random sampling, simple random sampling, systematic random sampling or cluster random sampling (Cohen et al., 2013). Non-probabilistic methods do not rely on random selection for reasons either of convenience or of research limitations. Certain sampling methods are bound to be more effective in certain cases and each method clearly has its own advantages and disadvantages. For instance, even a quantitative study might prefer employing a non-probabilistic method such as snowball sampling to probabilistic methods in a situation with a unique population, where its members are difficult to identify or very few in number (Cohen et al., 2013).

A first challenge for the research was obtaining access to a research population. This was difficult since a potential university would have to not only allow access to student data, but also aid in testing the software (e.g. by allowing students to use the software and sending them their performance predictions and software-generated advice). Inquiries were made to several universities in Saudi Arabia such as KFU, but the only one which agreed to fully assist with the research process was the University of Dammam. Access to such a small population and other factors such as incomplete data across all courses made selecting a probabilistic sampling method ineffective. However, while quantitative, probabilistic sampling is not absolutely necessary for this study, which analyses a prototype in an emerging field, and is in large measure still a work in progress.

The method chosen is a non-probabilistic sampling method through a case study approach. The study had access to student performance data across three courses but chose only one to conduct the research on since it had a more complete set of data, with many students being actively engaged in the course. 100 students, divided into two classes, were selected for developing the prediction model, taking the courses: Algorithm Analysis, Introduction to Algorithms, Data Structures, Object Oriented Programming, Advanced Algorithms, and Introduction to Artificial Intelligence. Their historic performance data was analysed through the use of data mining techniques such as Hidden Markov Models.

After the model was developed, two groups of 50 students each were selected through simple random sampling. Group A would make use of software recommendations, while Group B would not.

3.6 Research purpose

Noor (2008) identifies and classifies research by purpose into three possible areas: exploratory, descriptive and explanatory research. This study is firstly an exploratory investigation, which is done where a situation or problem is not clearly defined (Noor, 2008). In this study, there is indeed little past research into the application of data mining in E-learning, as is the application of advisory software based on statistical performance data. As such, the exploratory nature of this study aims to enable achieving a better comprehension of what the issues and challenges are in this area, suggesting further research. As a result of this, the study will be conducted as a case study, analysing student performance in a native environment. This is particularly important for studies where there is little previous information available, and where the unique nature of the situation requires use of a real setting in order to improve research accuracy.

Secondly, the study also has a descriptive component, which aims to analyse and discuss the current situation of data mining in E-learning, together with the various procedures those techniques may include such as: Hidden Markov Models, clustering, adaptive E-learning or learning style modelling. A descriptive aspect is also important in regard to data analysis and result interpretation, which are particularly critical for a quantitative study like the present one.

Beyond this classification, the study's purpose is to answer research questions, meet objectives and confirm or inform research hypotheses. The study will therefore look to improve comprehension of the research area, as well as achieve positive results in the specific research environment.

3.7 Data Collection and Analysis

Saunders and Lewis (2012) highlight the importance of well-chosen data collection and analysis methods, arguing they should flow naturally with the underlying assumptions of the research design. Also critical is ensuring that the data collected is accurate and that the analysis is effective and without bias.

Data was given to the researcher directly by the university, which also mediated the implementation of the software on student group A. Therefore, the researcher did not have direct contact with students throughout the research process. Saunders and Lewis (2012) also make a distinction regarding data collection processes, classifying them into mono, multiple, and mixed method data collection. This study utilises a mono method design. While beyond the scope of this research, a mixed method design could be used in future studies to gather qualitative data and gain insight into the student's perception of the recommendations and their effects on performance.

In building the prediction model, historical data of student performance at the end of semesters was obtained, which has been a partial limitation, as data was needed monthly, three times during each semester. This could negatively affect the performance of the prediction model.

3.7.1 Time Horizon

Another important aspect of the data collection process is its time horizon (Saunders and Lewis, 2012). Time horizon is the characteristic of a study, which describes the time period under which a phenomenon is researched: either a snapshot, as in cross-sectional research, or the phenomenon over a period of time, as in longitudinal research (Saunders and Lewis, 2012). The advantages of a longitudinal study are that it allows observing how a phenomenon unfolds and how variables change over time, thus having the potential to provide more insight. This study employs a longitudinal time horizon rather than a cross-sectional one, thus studying research questions across a period of time rather than in a specific moment (Saunders and Lewis, 2012). This is essential for this research because it allows analysing the performance of the SPS model in a real setting throughout time. A cross-sectional research would have imposed limitations by not allowing investigation into the performance of the advisory component.

Nonetheless, a longer period of time would have benefited the research by allowing this research to investigate the research questions more thoroughly, establishing averages in terms of performance prediction accuracy. Further, longitudinal research is also important in this area in order to show how performance prediction/advisor software improves over time, as more knowledge is accumulated.

3.7.2 Secondary Data

While primary data was critical in building the prediction model and conducting the experiment, the research also utilises an array of secondary data from books, journals, articles or case studies, which constituted the initial theoretical underpinning of this study. In particular, the secondary data consists of research into theoretical models, machine learning and data mining. These have been used as a starting point for building the algorithm and model for the SPS. Secondary data has advantages such as its ease of accessibility, which is particularly important for studies with low budgets (Saunders and Lewis, 2012). This study has made use of existing data on data mining techniques, E-learning tools, and other models of predicting educational performance. All research material has been referenced, and authors have been credited.

3.7.3 Data Analysis

Data analysis will be conducted in order to confirm or inform the research hypotheses, based on a framework of analysis. The first hypothesis regarding the ability of the SPS model to predict performance will be tested by comparing prediction accuracy to a neutral prediction such as assuming that the performance remains identical over time. If the results are positive, then the first hypothesis can be confirmed.

The second hypothesis (that the advisory component can lead to higher student performance) will be tested by comparing the level of performance improvement of Group A (which receives advice) to that of Group B (which does not receive advice). Positive results would confirm the second hypothesis. The standard deviation will also be considered in order to assess how advice affects low performance, average performance, and high performance students differently.

3.8 Reliability and Validity

A critical aspect of any research, particularly important to quantitative studies, is reliability and validity (Bush, 2007). Reliability deals with the measurement procedure and its ability to produce consistent results over time, enabling research to be generalised and applicable to a larger group. Validity on the other hand deals with the intrinsic logic of the research, assessing whether the research does indeed study what it intends to study.

Oluwatayo (2012) highlights the difficulties of achieving reliable and valid research studies in educational research, because of its intrinsic complexities. Oluwatayo notes the vast differences across student learning methods, personalities and cultures. Also noted are challenges of conducting the research in live settings without influencing participants (Oluwatayo, 2012).

Assessing reliability is closely linked with the sampling choices made by the researcher. Due to limitations forcing the use of a non-probabilistic small sample, the final results may have a low level of reliability. To be considered are also reliability limitations in a wider context of educational research:

- There are likely differences between individual students in the same course, between students in different courses and students in different universities. Social environment and social class background could also have an impact.

- The country of the university and its cultural background may also impact research results. For instance, studies such as those of Hofstede (2010) identify important cultural differences which may impact educational processes (e.g. culturally-specific uncertainty avoidance or long term orientation). In this sense, future development in E-learning must consider cultural background, which is particularly important in countries like the UK, where there is large diversity across the population.
- The field of study could impact results, which is particularly important since E-learning is not adequate for many subjects (e.g. medicine where practical application of knowledge is critical)

The study will also assess the validity of the two primary stages: developing the prediction model and conducting the experiment. Validity of the first stage will be impacted by reliability and the small sample size. Nonetheless, the prediction model has shown an approximately 80% accuracy in predicting student performance or aspects such as learning style, which is a good range in most Hidden Markov models (Van den Bosch, 2011). Also important is to note that accuracy is significantly higher than the accuracy of a neutral assumption (e.g. assuming student performance stays the same over time).

When discussing validity, it is also important to distinguish between internal validity and external validity (Onwuegbuzie, 2000). Internal validity investigates whether the results of the study are indeed only caused by the manipulated variable, while external validity assesses whether the results can be generalised to other research settings (Onwuegbuzie, 2000). External validity is easily ensured, as the prediction model can be applied to other sets of data, as well as other E-learning environments, given that these environments share common characteristics of data collection (are based on a common framework). Internal validity however is a more complex issue: the research design can clearly prove that personalised advice leads to better student performance. However, a question that arises is whether the statistical underpinning of the advice has any influence on student performance. Essentially, does it matter what exercises are suggested, or does it not given that they result in the same amount of time of additional work for the student? This could be looked into by further research that would study the effects of various types of personalised advice, leading to a better understanding of the best method of improving student performance.

3.9 Research Limitations

Flick (2015) highlights the importance of accurately identifying research limitations. According to Flick, research limitations are critical not just for identifying potential faults or insufficiencies of the research, but also for placing the research in the wider context of academic literature, and allowing researchers to solve limitations in further research.

There are several research limitations that influence the scope or accuracy of the final results. First of all, a significant limitation lies in the case study sampling approach, which provides neither geographical or cultural variety, nor diversity across the data through multiple universities. A broader scope could have been achieved by testing the algorithm on multiple universities' data (which would demonstrate a pattern) and across multiple countries (which would shed light upon the cultural dependencies, if any, of the SPS software). This was not possible, however, because of the data access limitations and low budget of this research project. These issues would have to be investigated in future research projects. Another aspect which needs further study is how performance prediction capabilities vary across different courses, since this research only studies students of a particular course.

Another limitation is the unavailability of data throughout the semester, which would have allowed for a better prediction model; in fact, data from only the end of semesters was available, whilst the software was designed with monthly data availability in mind. Also to be considered are the limitations of implementing the application on only a particular E-learning environment: EMES, within the CENTRA system. While the research does shed light on the issues studied and confirms the usefulness of the SPS, variation of E-learning environments would add increased depth, allowing the study to gain more insight into how data mining, E-learning and the student advisor interact in influencing student performance.

Furthermore, a limitation lies in the relative lack of previously available research into the application of data mining in E-learning. Therefore while this study investigates a basic issue of whether an advisor program based on data mining can improve performance, it is critical to also investigate how the various aspects of the program or the E-learning environment affect student performance. Further research could look into identifying the most effective combinations, techniques and advice strategies that best improve student performance.

Another limitation has to do with the unavailability of direct data, gathered from research participants. As such, all data was provided by the university, whose results or data collection process are unverifiable.

There is also a limitation regarding internal validity, where it is not clear how varying the content of software advice influences the level of achieved student performance. Most of these limitations can be resolved in further research. The reasons behind them have to do with the limited scope of this research: as the majority of the work was put into researching and building the actual software, and data access was severely limited, the main objective of the research is a preliminary demonstration of software capabilities. The research does not therefore aim to be exhaustive, which is difficult and will likely be achieved as the field of E-learning develops and matures within our society.

3.10 Ethical considerations

One of the most important aspects of research is ensuring ethical standards are properly respected (Saunders and Lewis, 2012). This should be shown not only in the treatment of research participants but also in the processes of data collection and data analysis. Participants must therefore be properly informed regarding the research, must give their consent and have their confidential information protected at all times (Saunders and Lewis, 2012). The data itself must be collected in an accurate and ethical way, without embellishing or cherry-picking, and must be analysed within an objective framework that limits researcher bias. Also essential are avoiding academic misconduct and ensuring all data from external sources is adequately referenced.

This research was designed as to abide by the guidelines of the British Educational Research Association (BERA, 2004). At the same time, the approval of the Research Ethics Committee was obtained, which allowed conducting the fieldwork in Saudi Arabia.

Throughout the research, contact with the student research participants has been indirect, mediated by the university. As such, the university is the only one that directly interacted with the students. However, sensitive data such as students' name and performance has been available to this study, which could raise ethical concerns if managed improperly. No personal identity data or individual performance data are therefore released in the presentation of results, but rather all data is presented as aggregate. All personal data has also been permanently shredded when no longer necessary, after the results of the second semester.

Additionally, in discussing the research process, the university has informed the researcher that all student participants have been assured regarding the confidentiality of their data. Students had also been informed about the right not to take part in the research or exit the programme at any time. Ethical aspects are also important concerning researcher objectivity and interpretation of results. In this regard, the researcher is particularly interested in establishing a clear framework of data analysis, which will define the terms under which hypotheses are validated or informed, leaving little room for researcher bias. These aspects are clearly detailed in the "3.7.3 Data Analysis" section.

Ethical issues of using a control group were taken into consideration while conducting this research. The debate within this area claims that regardless of whether the samples are in the control group or the experimental group, the ethical problems are the same: is it ethical to provide some students with an effective environmental intervention that other subjects do not receive (Bryman, 2012; Conner 1980). In this study, all participants were aware of the experiment that was being conducted. Consent was given to the researcher by students to conduct controlled group experiments; excluding the use of the SPS software, all students were receiving the same educational treatment. More importantly, the use of control groups was necessary to measure the change it will make if applied in the future.

3.11 Conclusions and Summary

Proper research design is essential in achieving a high level of research validity as well as reaching interesting results, of value to the specific field and research questions. This chapter has presented the main characteristics of the research design, describing its objectives, strategy, philosophy, sampling process as well as data collection and analysis. First of all, research objectives and questions have unveiled several key areas and outcomes for this research: building background in data mining tools and methods, building background in popular E-learning systems, designing and testing the SPS model in both prediction and advisory capabilities, evaluating results and discussing implications for the field of E-learning.

Research procedure and strategy were outlined in the context of research importance: improving E-learning procedures by better understanding of student performance. This would lead to better instructor capabilities as well as better student use of E-learning systems in a time where E-learning is more and more used because of distance and time limitations. The research procedures consist of several steps: researching data mining techniques, building the software prototype using Java, inputting historical data from UOD to build the prediction model, testing the prototype during the course of a semester and finally analysing results and reaching conclusions.

Research philosophy was described, and several key aspects were identified. The research thus employs an ontological assumption of objectivism and an epistemological assumption of positivism. The use of a deductive approach is also identified, with particular hypotheses that are made at the beginning of the research process.

Particularly important is the sampling process, which because of access limitations and the early nature of the field is a non-probabilistic case study approach. 100 students of a course at the University of Dammam were selected to build the prediction model. Out of these 100, two groups of 50 students were also selected by simple random sampling in order to assist with the experiment. An exploratory as well as descriptive research purpose has been identified, which enables the study to design data collection and analysis procedures. These procedures use a longitudinal time horizon, conduct statistical data analysis and also make use of secondary data. Research data is provided by the University of Dammam, with indirect contact with students in the researcher's case.

Key aspects of reliability and validity are then identified, with several reliability and validity issues present. In terms of reliability, there are several issues stemming from the small sample and limited availability of data such as not being able to ascertain cultural influences, influences of the particular course, field of study, or social environment and background. Validity is discussed through internal and external validity, which are primarily valid, with few additional issues. Research limitations however include the unavailability of complete data throughout the semester, little available research in this field or the inability to study certain aspects in depth. Ethical considerations were finally addressed, highlighting the nature of sensitive participant data, informed consent as well as ensuring confidentiality. Also discussed was researcher objectivity and how results interpretation must avoid bias.

Finally, the chapter has aimed to place the research into a wider context of the academic literature by highlighting its purpose, its limitations and its methods. Enriching academic literature and making research available for further investigation is the definitive purpose of the study.

Chapter 4 System Design

4.1 Introduction

This chapter deals with the proposed approach, the SPS system, developed with the purpose of predicting student performance. The overall design of the system and the components that comprise it are described. There are three main components:

1. ID3-based component
2. Hidden Markov Model (HMM) component
3. Ontology-based component

For each component, the motivation is discussed. It is very important to understand why the scholar had chosen to work with these components, based on the three algorithms. Then, there is an explanation for each component in detail. Finally, for each component, the researcher present practical examples that are intended to show that each component was successfully used. The results of using each component are presented with discussion and interpretation based on the outcome. By examples one can also understand how it will be used in real life situations.

The SPS system is a strong advisor, and will benefit both the students and their tutors. The unique design of using three different components is an attempt to increase the effectiveness of the final system. The adopted approach is based on three main components, each of them covers a different side of the problem. Merging the results of these three components gives the final prediction and detailed advice.

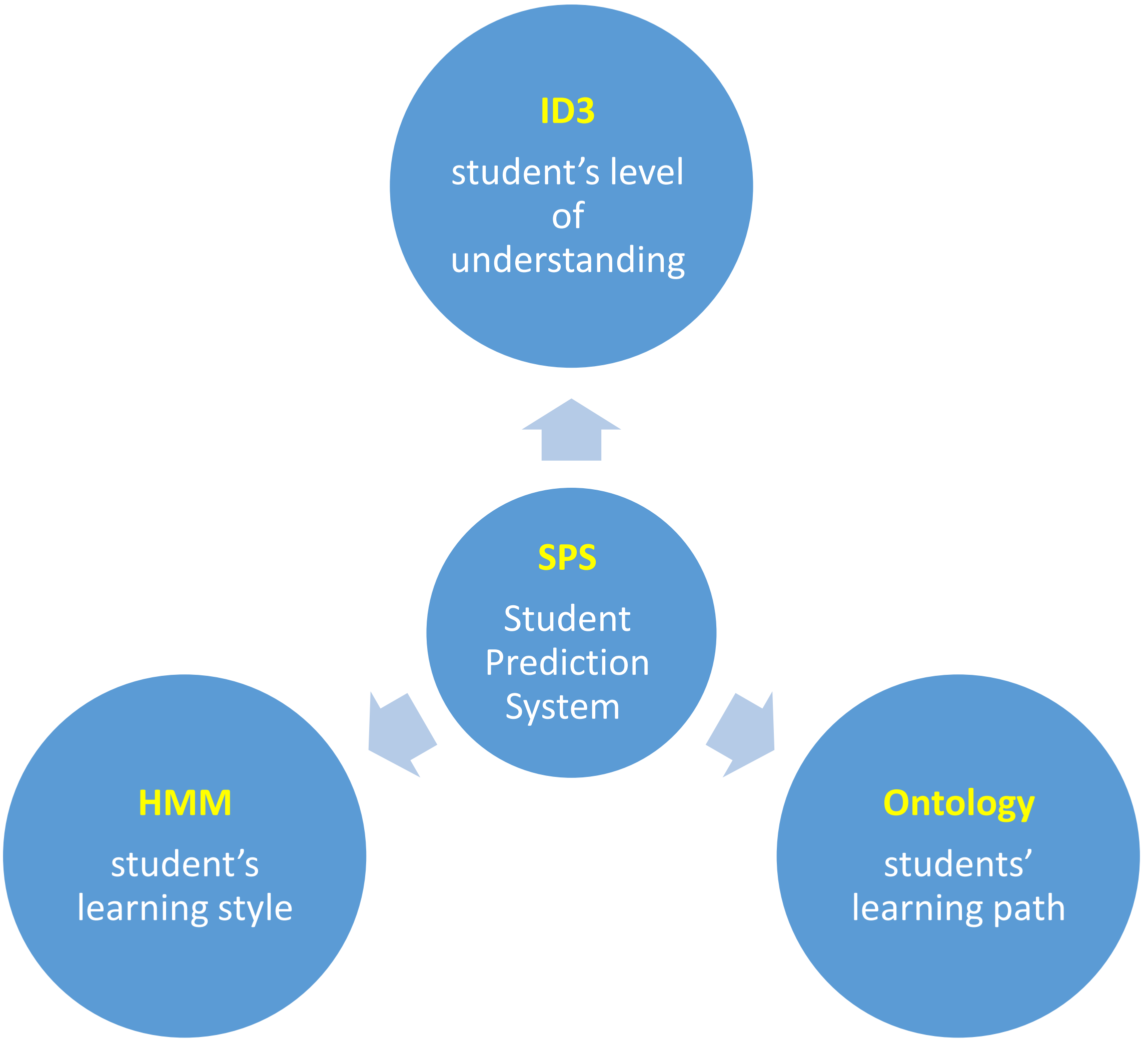


Figure 4 The three main components used

4.2 General Problems

To predict student performance using adequate models and help them in their studies, the general problems in this area must be understood. It is important to acknowledge previous approaches and the prediction models used. Various studies reveal sides to this problem that can increase the efficiency of the current approach.

Multiple factors can affect student performance. Understanding and measuring performance based on these potential factors is the only way to obtain models and approaches. Modelling must begin with data collection. Once data is at hand, it has to take a universal form to make it readable and usable. Well-defined criteria are set to find relevant values as to how close or far a student is from an ‘ideal’ performance. The ‘ideal’ only exists in theory. Performance, in practice, depends on how the researcher measured it.

Research in this field has accurately measured and defined certain factors that prove to be relevant to performance prediction. It is good to observe that different prediction models improved the performance of previous models. The approach presented in this thesis must also improve the performance and suitability of the already known results.

According to the study of Ramaswami and Rathinasabapathy (2012), a student’s academic performance is not easy to measure. The researcher must take certain factors into account to get an accurate result. Choosing a prediction model is only possible once the socio-economic, academic, and demographic factors are considered. These factors are deemed necessary because they will determine the performance of each student. Socio-economic factors will always influence performance. Areas where students are poor might either have no possibility to study enough or – on the contrary – have much more motivation in comparison with students from rich families. Demographic factors also influence the outcome. There are certain cases and areas of study where female students are more inclined to learn than males and vice versa. Lastly, the academic factors – such as how well-trained the tutors are, how good the library is – will influence the essence of the performance. If the tutors know their subjects very well, there is more probability of higher student performance than it is with low-performing tutors who are not proficient in what they teach. As the study reveals, the correct prediction algorithm’s model cannot be built simply, as it is the result of a complex process.

The mentioned scholars suggest using Bayesian Networks (BN) for the prediction. Bayesian Networks are a powerful choice thanks to the possibility to represent both elementary and complex problems in a simple manner. This representation requires to know how different features of the problem depend on each other. Marks obtained in secondary higher education were at the basis of the BN performance measurement. The scholars have used data about more than 5000 students, and beyond the mentioned socio-economic, academic, and demographic factors they relied on the grades. Thus, the BN is first trained and then they measured the accuracy of various search algorithms used, namely to assess how accurate the performance prediction was. Grades were also categorized differently in 2, 3, 5 and 7 categories, respectively. The scholars concluded that TAN was overall the most accurate algorithm in performance prediction. Advantages observed by the BN model include:

understanding and confirmation, earlier knowledge is included. Earlier knowledge is represented by information at hand, such as the student data collected in the mentioned paper. Understanding is achieved when observing the BN model nodes and arcs between the nodes. Each node represents a factor and each arc a relationship between two factors. Thus one can understand the BN model representation by understanding what relationships exist between different factors of the model. Confirmation is upon measuring the accuracy after training the BN model. The closer the outcome is to the expected, the better the model is confirmed to represent the problem correctly, namely by including necessary factors and the required relationships between these, which allow the model to produce correct results.

Ramaswami and Bashkaran (2010) attempted to collect data from students to analyse their performance. They have used a CHAID prediction model (Chi-squared Automatic Interaction Detection – decision tree technique first proposed in the 1980 PhD thesis of Gordon V. Kass) for the study. It was necessary to collect the survey answers and then standardise the given answers. According to a comparison with other models available at the time, the new results were classified as adequate. Finding the factors of true influence was a challenge. Their research relies on exactly 35 factors, from which the researcher mentions the student's family size, community, gender, and mode of transportation to school. The modelling also aimed to identify who the slower learners were.

Thai-Nghe et.al. (2010) attempted a newer approach, based on recommender systems. While people use the technique widely in e-commerce, e-Learning, and others, it proved successful in predicting student performance. To assess whether the approach is useful or not, they compared their recommender system prediction with linear regression and logistic regression using educational data. The study concluded that combining forecasting and factorisation (the proposed method in the paper) brings improvement when compared to the results of other scholars.

Another common problem is defining differences between slow and fast learners. As an attempt to solve the issue, Bhardwaj and Pal (2011) used a predictive data mining model. The aim was to process the collected data and use it with Bayes classification. The research makes it possible to identify performing students, who are fast learners. As the researcher identify the fast learners, slow learners study more with their teachers to achieve better performance. The scholars studied the effect of different factors on student performance. They analysed where the student lives, the habits, marks obtained during senior secondary education, the family's earnings and position, and a few more. The conclusion was that a desire to learn better is not the only influence on the performance, as other factors – from the ones mentioned – will also distinguish between fast and slow learners.

4.3 Using ID3 to predict student's level of understanding

4.3.1 Introduction

This section deals with a review of previous works that have used the ID3 algorithm to measure students' level of understanding. Various articles referenced in this section reveal the different approaches and results that have been obtained with ID3 decision trees and systems that have been implemented based on this widely used algorithm. The section is concluded by proposing a modified ID3 algorithm.

4.3.2 Review of previous works

There are several approaches to using ID3 to measure a student's level of understanding. It is fair to say that the level of understanding is in correlation with the expected grades. A student who does not understand matter A will not be able to study the subject with excellence. Their expected results are either failure or a low grade. If a student perfectly understands matter A, it is natural to say they will obtain very good or excellent results.

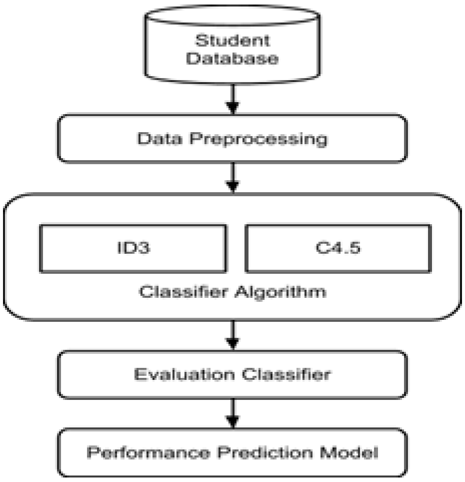


Figure 5 Representation of processing model used by Adhatrao et.al. (2013)

Adhatrao et. al. (2013) studied the potential of the ID3 and C4.5 algorithms. Their conclusion was that it is ideal to have a classifier algorithm that uses both ID3 and C4.5 for accurate results. The study was able to determine how freshly enrolled students are going to perform, based on how they studied before entering university. Thus, it was easy to see two main groups:

1. Students who had potential to excellence from the beginning
2. Students who need more help to develop

The overall process strongly relies on how information is gathered and what is included in the database. The correctness of the prediction based on the processing model from Fig.5 is between 75% and 78%.

Ogunde & Ajibade (2014) used ID3 to predict the student's level of understanding based on data collected upon admittance. The goal is to improve academic performance based on data that is readily available. Students fill forms, and databases load up with information from every student. It is important to understand the value of such information. It is important to use systems like the one presented, which rely on simple ID3 decision trees, correctly.

Maximising performance is crucial; which means that the depth of the decision tree should be reduced sufficiently to avoid overfitting. To achieve depth minimisation, the researcher know that every level represents a new decision based on some well-established rules. Based on the ID3 decision architecture, the researcher eliminated unnecessary decisions. To give an obvious example, predicting future grades based on height or weight is not possible.

Further performance improvement is also possible. In 2015, Joseph & Devadas modified the ID3 algorithm. The purpose of the approach is to get higher accuracy when predicting a student's performance based on prior results. Weight values are generated so that certain parameters will become more important. The difference from ID3 is using Gain Ratio and not Information Gain.

The accuracy of the prediction as measured by Joseph & Devadas in 2015 is in the table below.

Table 3 Accuracy measured in research by Joseph & Devadas (2015)

Algorithm	Accuracy of the prediction
C4.5	45.8333%
ID3	52.0833%
CART	56.25%
Modified ID3	76%

4.3.3 Motivation

ID3 is an algorithm that creates decision trees. Upon reviewing the literature, the researcher found that scholars have successfully used ID3 in their research, allowing them to improve performance, stability and accuracy.

Joseph & Devadas in 2015 proposed a new solution. While it is known that C4.5 is an effective algorithm, meant to improve ID3 and correct its weakness, the article referenced here proved a better approach. As it is known that C4.5 is a development of ID3, it was decided to use ID3 as a simpler approach for generating decision trees.

4.3.4 Discussion

The figure below shows the two phases of the Student Performance Prediction and Advising System (SPS). The first phase of the Performance Prediction Model Creator employs the utilization of specific courses by students to generate a performance prediction model. Importantly, course data is combined from previous semesters. This course data is retained in the database of the E-learning management system (LCMS). In the second phase, the generated performance prediction model is employed in the Performance Prediction Model Recommender to provide both instructors and students guidance regarding the overall performance of students benefiting from the E-learning system. A prediction model is achieved by the SPS through the use of a decision tree model.

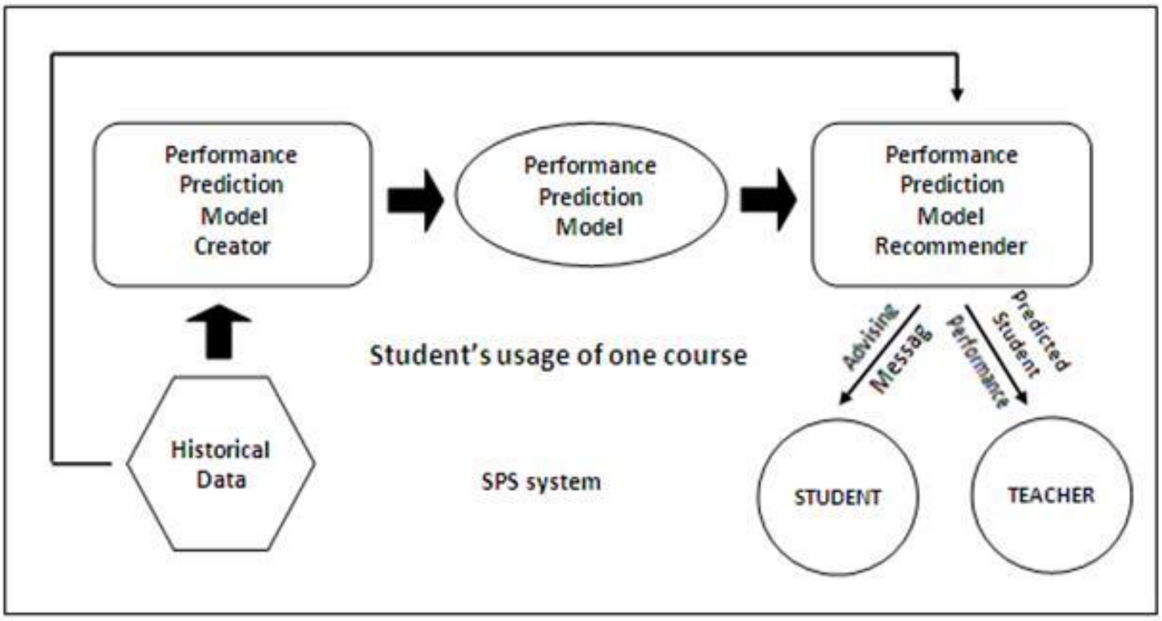

Performance Prediction Model Creator
Performance Prediction Model
Performance Prediction Model Recommender
Student's usage of one course
Advising Messag
Predicted Student Performance
Historical Data
SPS system
STUDENT
TEACHER

Figure 6 The SPS System

A student's performance is predicted by SPS by tracking from the root to a leaf of the generated decision tree prediction model. Whatever leaf is reached represents the value of the predicted performance of the student. Next, the student can utilize SPS to explain the basis of this predicted performance. Specific attributes of the students that led through this route on the decision tree will by explained here. Subsequently, the student will achieve subpar performance scores if he or she employs the attributes that were detected by the decision tree. On the other hand, if the student employs the attributes that would have resulted in a higher performance, he or she will achieve good performance results. This student clearly demonstrates to the student the rationale for the generated results.

Based upon the aforementioned results, SPS provides students with advice regarding ways in which they can improve their performance over the next month. This advice is generated by reverse tracking the decision tree prediction model for the following month. Here, the user chooses leaves representing higher performances and tracks them down to the originating root. By reverse tracking, the process identifies attributes that have required greater utilization than completed by the student and compares the necessary usage with that of the student. As a result, the student now has knowledge of how to increase performance over the next month.

The following examples reveal the ways in which students can take advantage of the results generated by the system. The employed data-mining algorithm depends on the following parameters: data from quizzes, data from assignments, and usage of the E-learning system. The following examples hypothesize the utilization of the selected attributes by the students. The utilization is employed to track the appropriate decision tree prediction model to achieve the desired performance.

4.3.5 Example

Let **S1** be a student with the following utilization scores in the **second month**:

1. *Number of completed quizzes:* **2 quizzes**
2. *Number of completed assignments:* **2 assignments**
3. *Number of pages read:***19 pages**
4. *Number of visits to LCMS:* **12 visits**
5. *Total time spent in LCMS in Hours:* **14 hours**

The decision tree prediction model for the second month is generated based on data from the second month of previous semesters. This data is mined and used in the case at hand, as shown in the figure below. By tracking this model from the root to a leaf by using the data from S1, the student will be able to visualize his or her historic performance during this time. In Example 4.5.5, the attained leaf is low thereby predicting a low performance of Student S1.

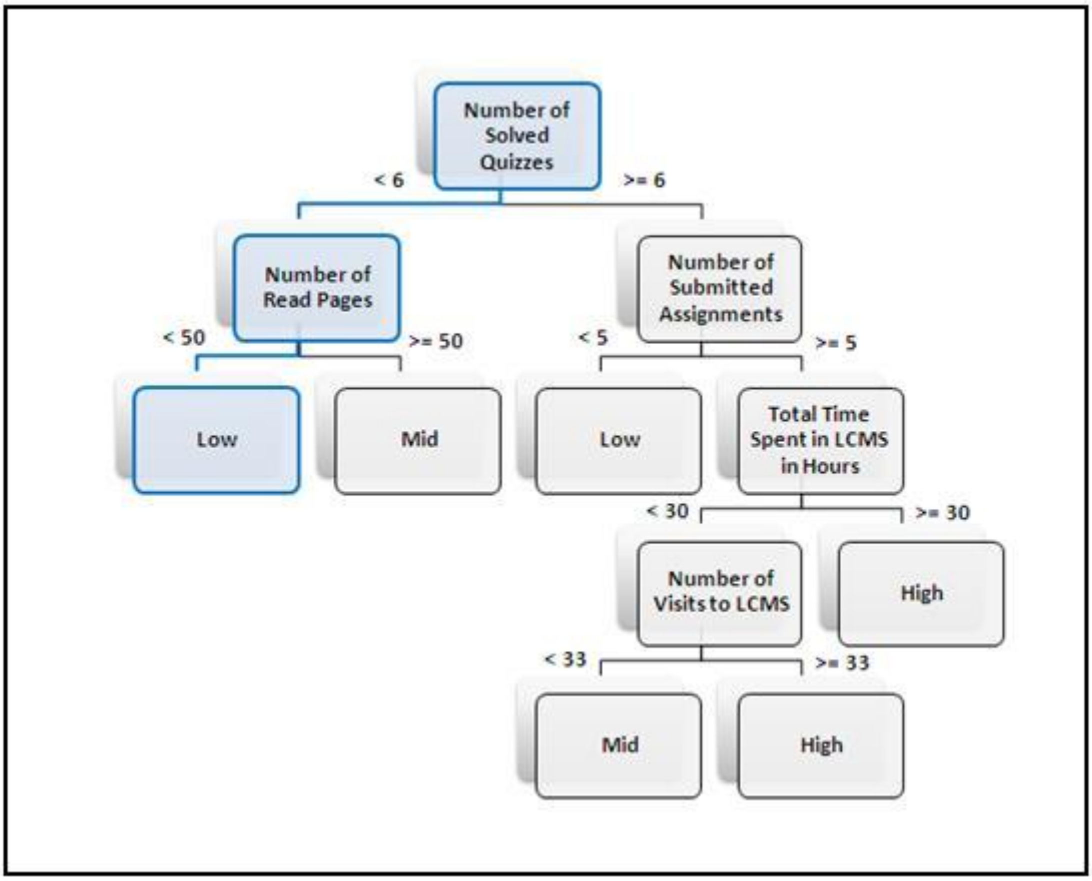


Figure 7 ID3 Example nr. 1

The advisory message to Student S1 is generated by storing data for student S1 regarding utilization of each parameter that will lead to attaining the "successful" side of the decision tree prediction model Number of read pages = 19 pages, and Number of solved quizzes = 2 quizzes. In addition, it stores the necessary utilization of these parameters: Number of read pages must be at least 50 pages, and Number of solved quizzes must be at least 6 quizzes.

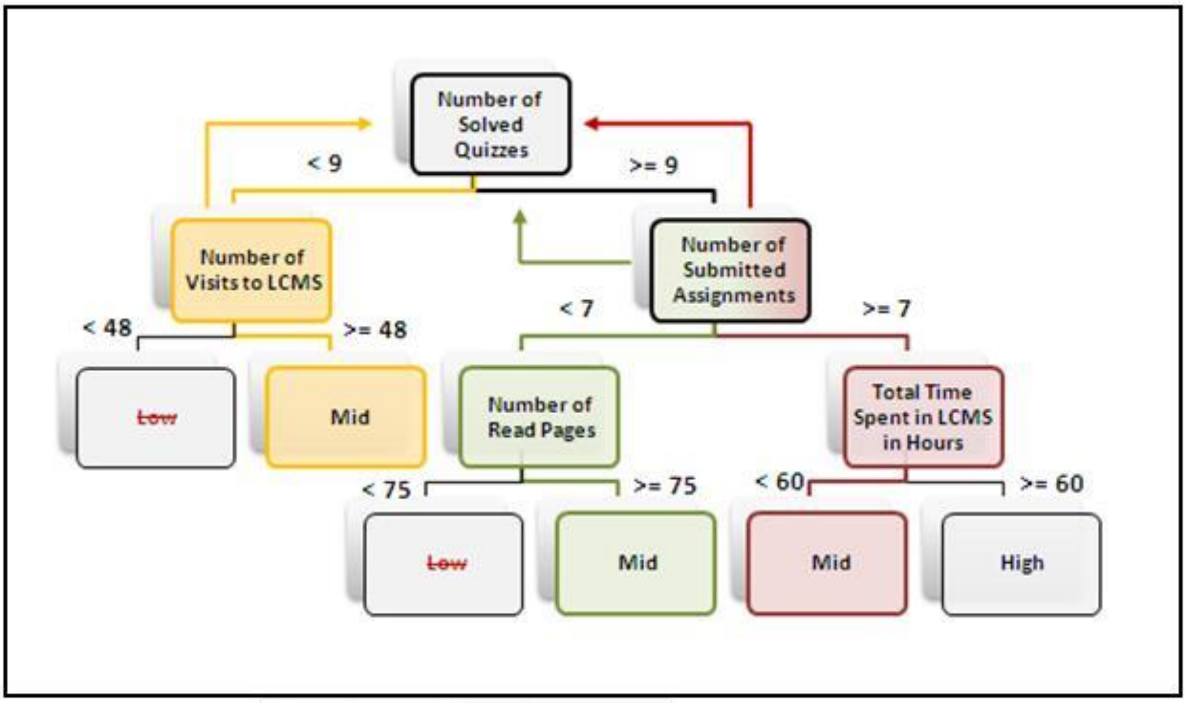


Figure 8 ID3 Example nr. 2

Consequently, Student S1 was sent a message showing a Low predicted performance, as well as his utilization of the detected parameters: Number of read pages, and Number of solved quizzes, an encouraging message expressing that the student needs to to work harder to achieve a higher performance rating, and advice on ways to achieve a higher performance rating in the next month.

Thus, the message of advice generated for Student S1 will look similar to below:

Number of solved quizzes = 2 quizzes → must be at least: 6 quizzes
OR
Number of read pages = 19 pages → must be at least: 50 pages

Taking user friendliness into account, the aforementioned example should alter its wording to the following:

- "You have solved 2 quizzes. Good job! We recommend solving at least 6 quizzes."
- "You have read 19 pages. Good job! We recommend reading at least 50 pages."

Next is the example of Student S2 whom has achieved the following parameter utilizations in the second month:

1. *Number of completed quizzes:* **5 quizzes**
2. *Number of completed assignments:* **4 assignments**
3. *Number of read pages:* **55 pages**
4. *Number of visits to LCMS:* **25 visits**
5. *Total time spent in LCMS in Hours:* **23 hours**

In this case, as like the first case, the decision tree prediction model for the second month is generated based on data from the second month of previous semesters. This data is mined and used in the case at hand, as shown in the figure below. By tracking this model from the root to a leaf by using the data from S2, the student will be able to visualize his or her historic performance during this time. In this example, the attained leaf is Mid thereby predicting a Mid performance of Student S2.

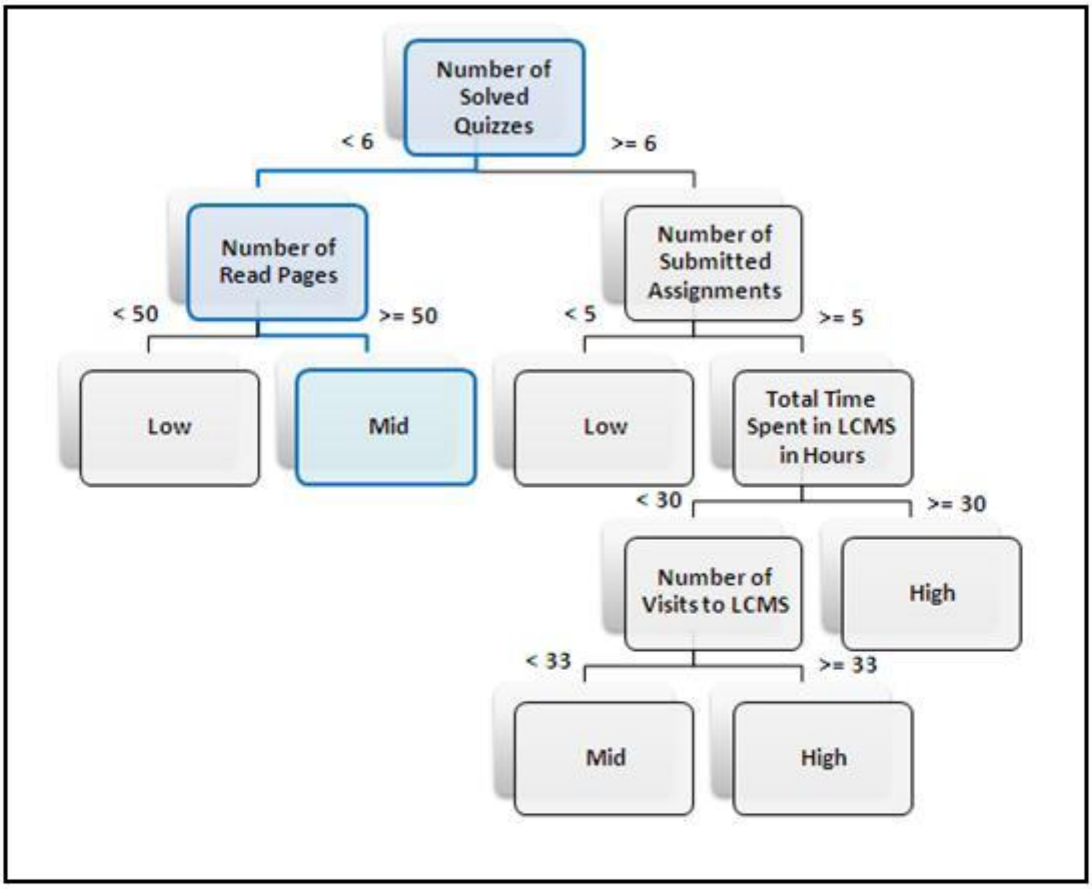


Figure 9 ID3 Example nr. 3

The advisory message to Student S2 is generated by storing data for student S2 regarding utilization of each parameter that leads to attaining the left side of the decision tree prediction model Number of solved quizzes = 5 quizzes. It also stores the utilization that should be attached in this parameter: Number of solved quizzes must be at least 6 quizzes. In addition, the SPS reverse tracks the decision tree prediction model of the third month to provide student S2 advice regarding ways to improve his performance to achieve a status of High in the third month.

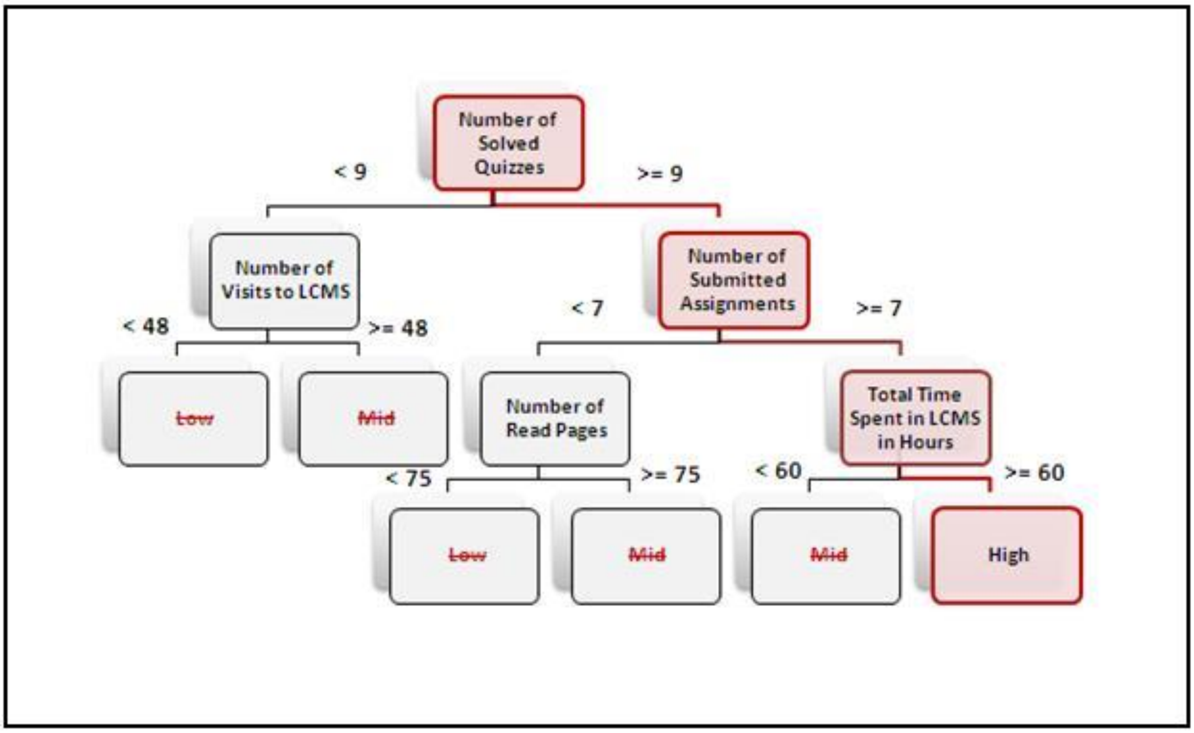


Figure 105 ID3 Example nr. 4

Accordingly, Student S2 was sent a message showing a Mid predicted performance, as well as his utilization of the detected parameters: Number of solved quizzes, an encouraging message noting to work harder to achieve a higher performance, and advice on ways to achieve a higher performance rating in the next month.

Thus, the message of advice generated for Student S2 will look similar to below:

Number of solved quizzes = 5 quizzes → must be at least: 6 quizzes

4.4 Using Hidden Markov Modelling (HMM) to recognize student’s learning style

4.4.1 Introduction

Researchers are created differently, with various opinions and views, in the same situation. The way researchers understand and see the world is very different, even our response to how researchers perceive information is quite different. Take, for example, students and the way they usually prefer to study for a lesson. Some of them choose to listen to a lesson that is mainly instructional, showing that the way they learn is through an auditory technique while some other students perceive the study materials as pictures, the so-called visual learner.

Some students can also have other characteristics when they study for a lesson such as a close physical interaction or contact with the material provided for learning. Such contact while studying is the character of the so-called tactile-kinesthetic learner. Likewise, other students tend to assimilate the information through making certain connections with their past or recent personal learning experiences. The latter is the so-called internal kinesthetic learner.

These types of characteristics about the user cognition are known as learning styles, but they have a wider range than the researcher can imagine. Researchers define a learning style as a conglomerate of affective, cognitive and psychological characteristic factors. These factors are relatively stable indicators of the way that learners perceive, then interact with and respond to the environment of learning.

In adaptive learning, the researcher consider that as an aiding navigator teacher or computer, where learning style is a major factor. Learning styles are analysed in a comprehensive manner in the theory of psychology, but there is a new approach to structure the learning styles based on mathematical tools that will predict or infer students' styles.

To understand how the Hidden Markov Models (HMMs) discover and represent the students' learning styles, it is important to first identify the way this approach functions. A well-known fact is that a HMM is a strong statistical tool. The HMM models generative sequences, which are the result of an underlying process that is able to produce a sequence that is measurable. There are the hidden and observable states of the model, and the way it works is intuitive. Based on known probabilities of going from state A to state B, the HMM generates certain sequences. The observable part is the sequence of symbols that will be generated, but that is based on the hidden state transitions. It is appropriate to use HMMs when the researcher want to predict the learning styles of students according to observed evidence gathered from them.

4.4.2 Review of previous works

Beal et al. (2007) modelled the learning patterns of high school students, and they used Hidden Markov Models. Hidden states are used to measure the engagement of individual students: low, medium or high. Transition matrices are important to understand the engagement of a student. If the student has a certain engagement at a given time T, the researcher can easily find the probable engagement at time T+1 by following the matrices.

Table 4 Transition matrices for CA and BA students by Beal et al. (2007)

Sample	Hidden state	Low	Medium	High
CA	Low	0.3014	0.2423	0.4563
CA	Medium	0.1982	0.4721	0.4176
CA	High	0.1517	0.3312	0.6050
BA	Low	0.4727	0.1966	0.3307
BA	Medium	0.1720	0.5583	0.2696
BA	High	0.1470	0.2772	0.5758

BA and CA denote the different students targeted in this research. BA were the 122 students who had 30 math problems solved (group average), while CA were the 91 students who had only 20 math problems solved (also group average). The values in the table represent the probability of a student from a certain group in a certain state to move to another state. For example, a medium CA student has a 19.82% chance of becoming a low CA student. As another example, a BA student with high level of engagement will have a 57.58% chance of remaining with the high level of engagement.

When measuring transitions for more students, it was observed that there are certain attitudes or future learning styles. In most cases, there is a steadiness in engagement. Steadiness means that low students remain low, medium remain medium and high remain high. A few students will get more or less engaged. It is remarkable that the likelihood to become low in engagement is only about 20% on the samples studied. When measuring on 91 CA students and 122 BA students, the researcher can observe 3 groups for the CA and 4 groups for the BA.

The observation was that the BA students weren't as engaged as the CA students. When measuring the prediction accuracy, the report compared HMMs to MCs (Markov Chains). It is obvious that HMMs can predict with higher accuracy as shown in the table below.

Table 5 HMM versus MC accuracy from the study of Bael et al. (2007)

	Ind. HMM	**Ind. MC**	**Gr.HMM**	**Gr. MC**
CA	42.13%	33.43%	34.40%	18.73%
BA	48.35%	32.48%	26.00%	15.52%

As a conclusion, HMMs prove to be efficient and able to predict variations that can occur in students' future engagement patterns.

Johns & Woolf (2007) came up with a new approach that presented a dynamic combination of two models, namely the Item Response Theory (IRT) and the HMM.

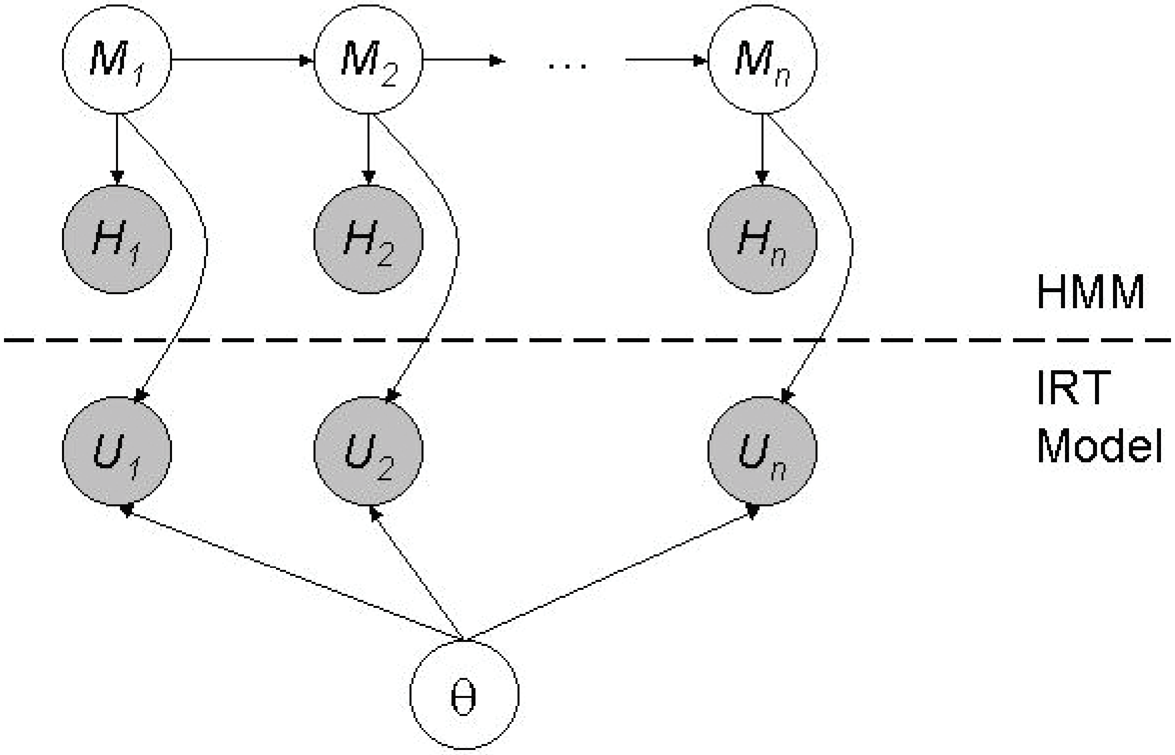


Figure 11 The Dynamic Mixture Model presented by Johns & Woolf in 2007

Student motivation, proficiency, evidence of motivation and response to problems are the four variables. Two latent variables are defined. Proficiency is denoted θ and motivation is denoted M_i. Two observed variables are also defined: U_i represents the initial response of a student, while H_i represents the hidden motivation variable. The possible values for U_i are correct or incorrect, while H_i will be 'many-hints', 'quick-guess' or 'normal', based on the way the study defines motivation levels: 'unmotivated-hint', 'unmotivated-guess' and 'motivated'.

The connection between these variables is the Dynamic Mixture Model (DMM), as it appears in Figure 12. To validate the results, they used cross validation. The resulting model's name is DMM-IRT, representing the dynamic combination.

The accuracy measurement results are presented in Table 7 below.

Table 6 Accuracy measurement presented in the research of Johns & Woolf (2007)

Model	**Cross Validation Accuracy**		
	Average	**Minimum**	**Maximum**
Default	62.5%	58.2%	67.7%
IRT	72.0%	70.4%	73.6%
DMM-IRT	72.5%	71.0%	74.0%

4.4.3 Motivation

In most cases the features which have been used in the last component could give an idea about the level of the student and give some general advice, but it does not reflect the main causes of a student's weakness. Therefore, the output of a system which bases on this set of features would be general. In this component, the researcher will consider the personality of the user to be more adaptive i.e. to give advice which considers a student's learning style. Many cognitive studies show that each person has a different learning style. Such research shows that when learners learn with their preferred learning style, they will take less time to master a given concept in a more comprehensive way. So, determining the learning style helps the system in finding the most suitable advice for each student.

Different learning style models have been proposed each of which is based on various aspects. “The most known model is VAK model, this model concentrates on the human observation channels; vision, hearing and feeling. Depending on this model, most learners can be categorized as Visual, Auditory or Kinesthetic (VAK) learners based on how they prefer to receive and process information” (Surjono, 2011, p.2351). Due to its simplicity, this is the most popular model nowadays. Visual learners have preferential learning in terms of seeing pictures, diagrams, slides, handouts, maps, tables, and charts. Auditory learners prefer to learn through listening to lectures, participating in discussions and talking about subjects, while kinesthetic learners prefer to learn by experience and experiments like moving, touching, exploring and experimenting (Surjono, 2011).

The VAK model was considered as it is widely used among scholars in the field and considered as one of the important models that explores students’ learning styles (Pashler et al., 2008; Brown et al., 2009).

The benefits of VAK learning styles are:

- Suitable for different types of learners which would help the researcher to cover all types of students in the class
- Encourage students to participate
- Make classes more interesting as it uses different ways to make the participants engage in the educational process.
- It has proved to be a successful method to teach individuals e.g. English language learning (Gilakjani, 2011). Moreover, a study revealed that using the Visual component of the VAK model enabled students to evaluate their teachers, which lead to better teaching (Hamdy, et al., 2001).
- Help in supporting students who have special needs or requirements.

Because of the previously mentioned advantages and wide use of VAK among British educational institutions the researcher adopted the VAK model as the primary technique to identify student learning styles. It is important to mention that other learning styles models such as Dunn & Dunn were considered (See 2.3.11.3 Learning Style Models) but VAK was thought the most appropriate model.

Hidden Markov Models will also help in making the system adaptive to users' personalities. Assuring the adaptive aspect is important, because every student will work more or less based on their personalities and learning styles.

4.4.4 Discussion

To consider the learning style in the system the researcher need to deal with a number of challenges:

Determine the learning style of each student

To do this the researcher needed to add more features which represent the styles. One of the ways which can be used is a questionnaire which contains several questions about learner personality, attitude and behaviour. Based on much cognitive research, the researcher designed this questionnaire which consists of 30 questions (See Appendix D), each of them has three different answers and each of them reflect certain learning styles.

Students should complete this questionnaire one time after the first usage of the system. For each question, the student should choose only one answer.

The 30 questions which are collected from online resources and their answers are as follows:

1. When I operate new equipment I generally:

a) Read the instructions first

b) Listen to an explanation from someone who has used it before

c) Go ahead and have a go, I can figure it out as I use it

2. When I need directions for travelling I usually:

a) Look at a map

b) Ask for spoken directions

c) Follow my nose and maybe use a compass

3. When I cook a new dish, I like to:

a) Follow a written recipe

b) Call a friend for an explanation

c) Follow my instincts, testing as I cook

4. If I am teaching someone something new, I tend to:

a) Write instructions down for them

b) Give them a verbal explanation

c) Demonstrate first and then let them have a go

5. I tend to say:

a) Watch how I do it

b) Listen to me explain

c) You have a go

6. During my free time I most enjoy:

a) Going to museums and galleries

b) Listening to music and talking to my friends

c) Playing sport or doing DIY

7. When I go shopping for clothes, I tend to:

a) Imagine what they would look like on

b) Discuss them with the shop staff

c) Try them on and test them out

8. When I am choosing a holiday I usually:

a) Read lots of brochures

b) Listen to recommendations from friends

c) Imagine what it would be like to be there

9. If I was buying a new car, I would:

a) Read reviews in newspapers and magazines

b) Discuss what I need with my friends

c) Test-drive lots of different types

10. When I am learning a new skill, I am most comfortable:

a) Watching what the teacher is doing

b) Talking through with the teacher exactly what I'm supposed to do

c) Giving it a try myself and work it out as I go

11. If I am choosing food off a menu, I tend to:

a) Imagine what the food will look like

b) Talk through the options in my head or with my partner

c) Imagine what the food will taste like

12. When I listen to a band, I can't help:

a) Watching the band members and other people in the audience

b) Listening to the lyrics and the beats

c) Moving in time with the music

13. When I concentrate, I most often:

a) Focus on the words or the pictures in front of me

b) Discuss the problem and the possible solutions in my head

c) Move around a lot, fiddle with pens and pencils and touch things

14. I choose household furnishings because I like:

a) Their colors and how they look

b) The descriptions the sales-people give me

c) Their textures and what it feels like to touch them

15. My first memory is of:

a) Looking at something

b) Being spoken to

c) Doing something

16. When I am anxious, I:

a) Visualize the worst-case scenarios

b) Talk over in my head what worries me most

c) Can't sit still, fiddle and move around constantly

17. I feel especially connected to other people because of:

a) How they look

b) What they say to me

c) How they make me feel

18. When I have to revise for an exam, I generally:

a) Write lots of revision notes and diagrams

b) Talk over my notes, alone or with other people *c) Imagine making the movement or creating the formula*

19. If I am explaining to someone I tend to:

a) Show them what I mean

b) Explain to them in different ways until they understand

c) Encourage them to try and talk them through my idea as they do it

20. I really love:

a) Watching films, photography, looking at art or people watching

b) Listening to music, the radio or talking to friends

c) Taking part in sporting activities, eating fine foods and wines or dancing

21. Most of my free time is spent:

a) Watching television

b) Talking to friends

c) Doing physical activity or making things

22. When I first contact a new person, I usually:

a) Arrange a face to face meeting

b) Talk to them on the telephone

c) Try to get together whilst doing something else, such as an activity or a meal

23. I first notice how people:

a) Look and dress

b) Sound and speak

c) Stand and move

24. If I am angry, I tend to:

a) Keep replaying in my mind what it is that has upset me

b) Raise my voice and tell people how I feel

c) Stamp about, slam doors and physically demonstrate my anger

25. I find it easiest to remember:

a) Faces

b) Names

c) Things I have done

26. I think that you can tell if someone is lying if:

a) They avoid looking at you

b) Their voices changes

c) They give me funny vibes

27. When I meet an old friend:

a) I say "it's great to see you!"

b) I say "it's great to hear from you!"

c) I give them a hug or a handshake

28. I remember things best by:

a) Writing notes or keeping printed details

b) Saying them aloud or repeating words and key points in my head

c) Doing and practicing the activity or imagining it being done

29. If I have to complain about faulty goods, I am most comfortable:

a) Writing a letter

b) Complaining over the phone

c) Taking the item back to the store or posting it to head office

30. I tend to say:

a) I see what you mean

b) I hear what you are saying

c) I know how you feel

The 30 questions have been selected carefully. Each of the questions with each of the possible answers covers certain aspects of the learning styles. It is possible that other questions could have been used and included, but upon developing the SPS system, these questions presented have been efficient in representing the learning styles of each student. the researcher continues the discussion to understand how these will be represented.

Representing the learning style of each student

After the questionnaire, the next challenge would be representing the learning style of the student. Theoretically, there are three learning styles: Visual, Auditory or Kinesthetic. Practically, students could not be categorized in one class. For example, no learner is 100% a visual learner. Most “visual learners” might be for instance 65% visual learners, 20% auditory learners, and 15% kinesthetic learners. Each learner possesses several learning styles which they can mix for obtaining the most convenient combination of learning activities. For this reason, the researcher represented the learning style of each student in a “fuzzy way”; practically, the researcher represented it as a vector of three percentages, one for each learning style. For example, if the learning style of a user is (75%,10%,15%) this means that they are a visual learner with 75% probability, an auditory learner with 10% probability, or a kinesthetic learner with 15% probability. Clearly the sum of these three probabilities should be 100%.

These three probabilities were calculated depending on students’ answers:

- The percentage of being a visual learner is the number of ‘A’ answers divided by the total number of questions.
- The percentage of being an auditory learner is the number of ‘B’ answers divided by the total number of questions.
- The percentage of being a kinesthetic learner is the number of ‘C’ answers divided by the total number of questions.

Depending on the last three values, each learner is classified into seven types (seven fuzzy learning styles) depending on the following rules:

- If V-probability (the probability of being a visual learner) is larger than 60%, the user is classified as class T1.
- If K-probability is larger than 60%, the user is classified as class T2.
- If A-probability is larger than 60%, the user is classified as class T3.
- If A-probability is larger than 40% and K-probability is larger than 40%, the user is classified as class T4.
- If A-probability is larger than 40% and V-probability is larger than 40%, the user is classified as class T5.
- If V-probability is larger than 40% and K-probability is larger than 40%, the user is classified as class T6.
- If the three probabilities are less than 40%, the user is classified as class T7.

Now that it is seen how the learning style is represented, an efficient model must be built, thus the choice of HMM. The researcher will now discuss in detail how the model is built based on the information already presented. The final purpose of the model will become the possibility to give quality advice to each student.

Using the Learning Style information to give more accurate prediction and advice

In this component, the researcher will use HMM to build this model. In the HMM system, the hidden states will be one of the three states: Visual (V), Auditory (K) and Kinesthetic (K).

The transition from one hidden state to the next will be captured using the state transition probabilities matrix. Formally, the state transition matrix is:

$$A = \{a_{ij}\} = \begin{array}{cc} & \begin{array}{ccc} V & A & K \end{array} \\ \begin{array}{c} V \\ A \\ K \end{array} & \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \end{array}$$

$$a_{ij} = P\big(state\ q_j\ at\ time\ (t+1) | state\ q_i\ at\ time\ t\big)$$

The Markov process is assumed to be of order one so it will depend on one preceding state. The observables in this case will be test results gathered for each student at the end of each E-learning session. The observables, for simplicity, will be categorised as "Improved Performance" (I or 0), "Static Performance" (S or 1) and "Deteriorated Performance" (D or 2).

The observation probability matrix is:

$$B = \{b_j(k)\} = \begin{array}{cc} & \begin{array}{ccc} I(0) & S(1) & D(2) \end{array} \\ \begin{array}{c} V \\ A \\ K \end{array} & \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \end{array}$$

$$b_j(k) = P(observation\ k\ at\ time\ t | state\ q_i\ at\ time\ t)$$

The observation probability matrix will be defined and improved upon by an iterative process. The initial state distribution matrix, π, can easily be defined using the clustering data obtained earlier.

Training the Model

In order to use the HMM for intelligent content, A and B must be computed. To do so, the HMM must be trained by using the training data. Training data will consist of an initial test student group. During the training of the HMM system, an observation sequence is obtained and is represented as:

$$\mathcal{O} = (\mathcal{O}_0, \mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \ldots\ldots, \mathcal{O}_{T-1}) = \ Observation\ Sequence$$

For example, $\mathcal{O}$ = (1, 0, 1, 1, 2, 0, 0, 1) is an observation sequence of length eight. The observations correspond to "Improved Performance" (0), "Static Performance" (1) or "Deteriorated Performance" (2).

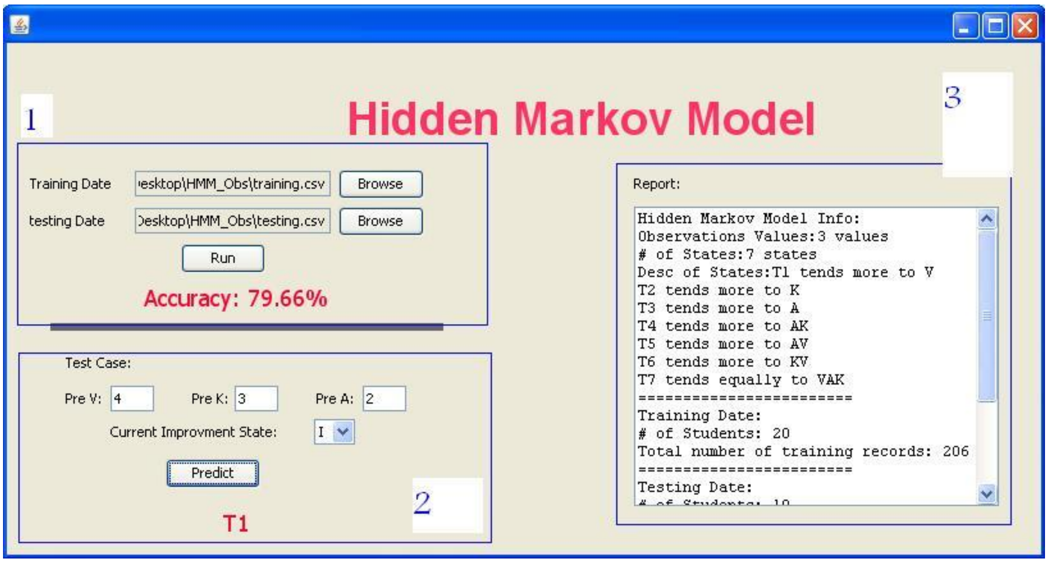
4.4.5 Example

To illustrate the proposed approach, an example is presented here. This example was done in three stages. The three stages are presented in the figure below (see Fig.12), where numbers 1, 2 and 3 respectively represent Part 1, Part 2 and Part 3 respectively. Each part means the following:

Part 1: two CSV files were entered into the system (from the collected data in a previous step), one for training and the other for testing. Then the system trains the HMM which has been decried in the design and then the researcher tested the trained HMM using the testing data and show the results.

Part 2: is just a simple UI to be able to do some fast tests for the system. This component is just for testing and providing simple demos (see Figure 12). It is important to remember that V,A,K refers to Visual Auditory Kinesthetic learning styles. The case in the image means: if V=4 (of ten i.e. 4/10), K=3, A=2 and the student is improved then the student tends to be Visual learner (the state T1). The learning style of the student is expected to be Visual. (in this stage the researcher will use this information to give more suitable suggestion to the learner regarding their learning practices)

Part 3: is a report for the training operation which describes the HMM items and its results. Again here V,A,K refers to Visual Auditory Kinesthetic learning styles.



Training Date: 'esktop\HMM_Obs\training.csv

testing Date: Desktop\HMM_Obs\testing.csv

Accuracy: 79.66%

Test Case:

Pre V: 4 Pre K: 3 Pre A: 2

Current Improvment State: I

T1

Report:

```
Hidden Markov Model Info:
Observations Values:3 values
# of States:7 states
Desc of States:T1 tends more to V
T2 tends more to K
T3 tends more to A
T4 tends more to AK
T5 tends more to AV
T6 tends more to KV
T7 tends equally to VAK
========================
Training Date:
# of Students: 20
Total number of training records: 206
========================
Testing Date:
# of Students: 10
```

Figure 12 Hidden Markov Model example

4.5 Using an Ontology in students' learning path

4.5.1 Introduction

In this section the researcher first introduces the concept of ontology. Then follows a discussion about why and how this relates to software design, and how a student model can be built. Then, a review of how previous works have used ontologies and built student models based on the ontology modelling of their choice is given.

According to the Merriam-Webster dictionary, ontology means two things:

1. a branch of metaphysics concerned with the nature and relations of being
2. a particular theory about the nature of being or the kinds of things that have existence

Ontology deals with the way that a student learning behaviour is described, and the relations between their particular characteristics. For example, a student can have a motivation which is determined by several factors or another example is that a student’s learning background can have a range of materials they learned over many years.

In general, ontologies refer to the nature of existence, and in software programming they have a particular meaning as the relationships or concepts existing for a particular agent or a particular community of agents. This definition follows the traditional use of ontology as a set of concepts but is more general.

While ontology is typically understood as a class hierarchy in a taxonomic sense, it can also be understood as a class definition or as simple subsumption relations. Nonetheless, ontologies do not have to be limited to such a strict form. Ontologies can also be understood outside conservative definitions that focus more on terminology and the traditional logic sense than improving knowledge regarding a subject.

4.5.2 Review of previous works

An ontology is a natural and effective way to describe a students' learning path. Paneva in 2006 used a simple student ontology as it appears in the figure below.

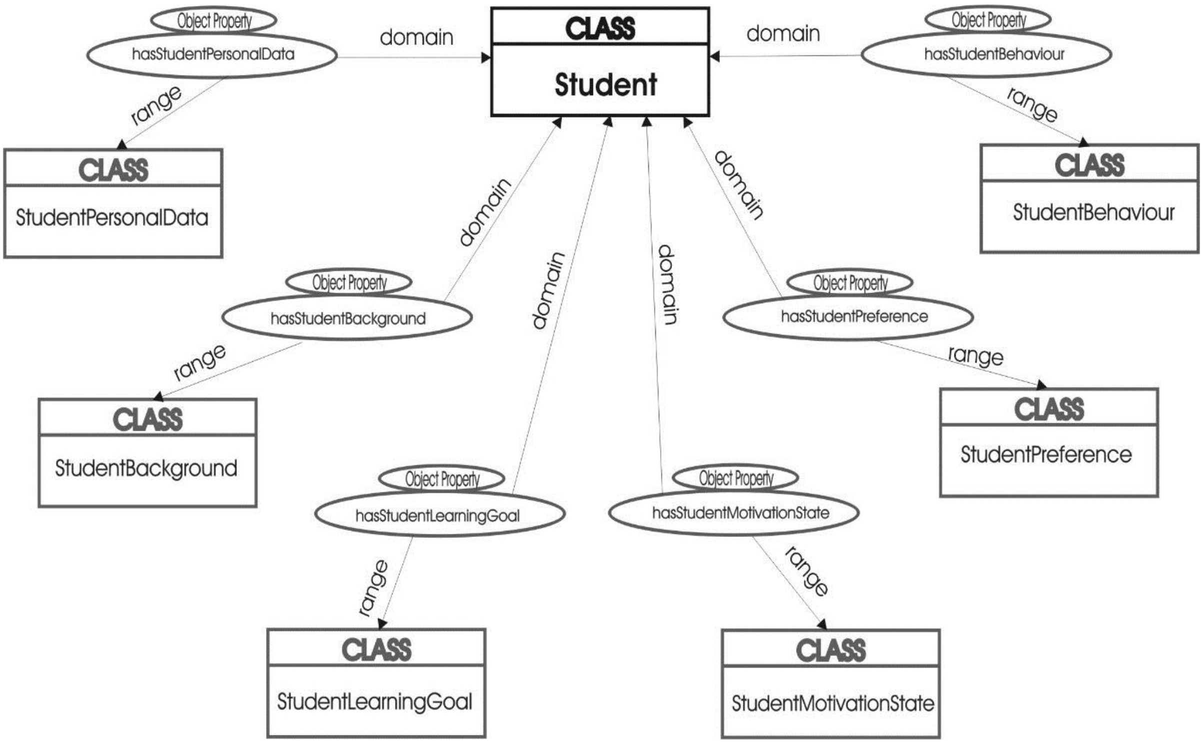


Figure 13 Student ontology as used by Paneva in 2006

Paneva wrote about using ontology-based students' learning paths. The research included the method to build an ontology-based students' learning path. In the article, OWL is used as the ontology language.

General information and the behaviour of each student are the two separate parts of the used ontology. In general information, they have included data about previous education, experience, motivation state, knowledge level, name, age, and more. Behaviour information can be competence level, object observation time, chosen object and so on. The model, based on the above decisions, can be used efficiently.

Furthermore, the OWL language needs object properties to be defined - relations between students. For this, the article shows how to smartly implement inverse properties. This means that if student A is linked to student B with a property, the inverse property of the same will link student B to student A. An example would be *hasInterest* and *isInterestOf*.

The ontology model presented by Paneva in 2006 is general, meaning it is not fully complete. A better ontology would be adaptive, so that it would adjust to changing student behaviour over time.

In 2009, Pramitasari et al. continued to research the idea of how an ontology can adapt itself to the student. They stated that for human teachers it is not always easy to consider many students, from which some have a certain learning process and the others have their own processes as well. A fair representation is to understand the distinction between visual and verbal learners.

Lately there is a shifting towards broader use of E-learning. The research paper focuses on using learning style and prior knowledge of students to create a new students' learning path, using an ontology. In other words, they personalised the E-learning system so that each student can learn with their own styles.

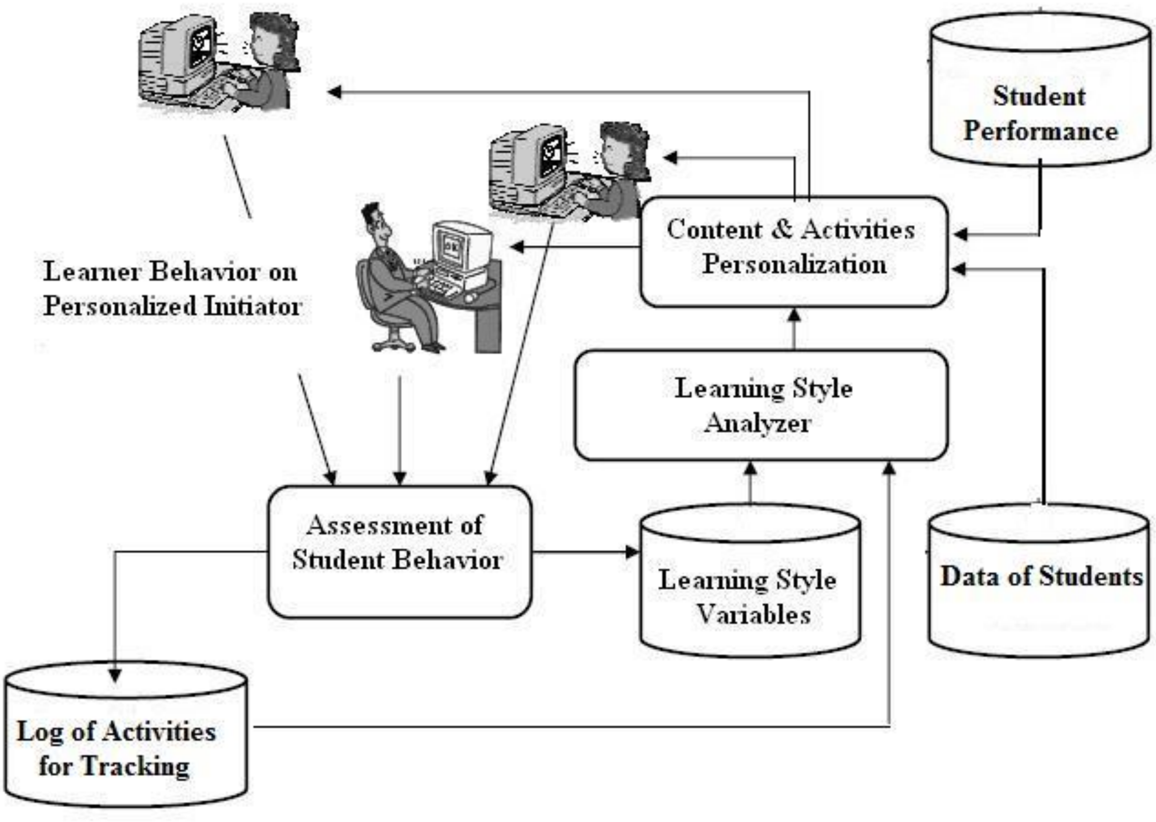

Student Performance
Content & Activities Personalization
Learner Behavior on Personalized Initiator
Learning Style Analyzer
Assessment of Student Behavior
Learning Style Variables
Data of Students
Log of Activities for Tracking

Figure 14 Personalization model framework by Pramitasari et al. (2009)

Figure 15 shows how the personalisation model works. Student behaviour is assessed, and then the learning style analyser will aid the content and activities personalisation, which also takes student data and performance into account. The research concluded that the created ontology with its rules was working properly, as was expected.

Day et al. developed an interesting concept in 2005. They used an ontology to develop an intelligent tutoring agent (ITA) which worked through IM (instant messaging). The goal of the research was creating an environment where the ITA would be an assistant for the human teacher, and also aid in the communication between students and teachers. The implementation only works with MSN Messenger, and only if the specified MSN ID is added. The ITA concept which uses the INFOMAP ontology is another side of using ontologies, but targets helping the human teacher and the student as well. INFOMAP was developed by Hsu et al. in 2001 and is a framework meant to represent knowledge, including domain knowledge, linguistic knowledge and common-sense knowledge.

Also in 2005, Brusilovsky et al. introduced a new concept called ADAPT^2. In the architecture of ADAPT^2, there are multiple components, but the researcher will focus on the Ontology Server. Below is a representation of the entire system:

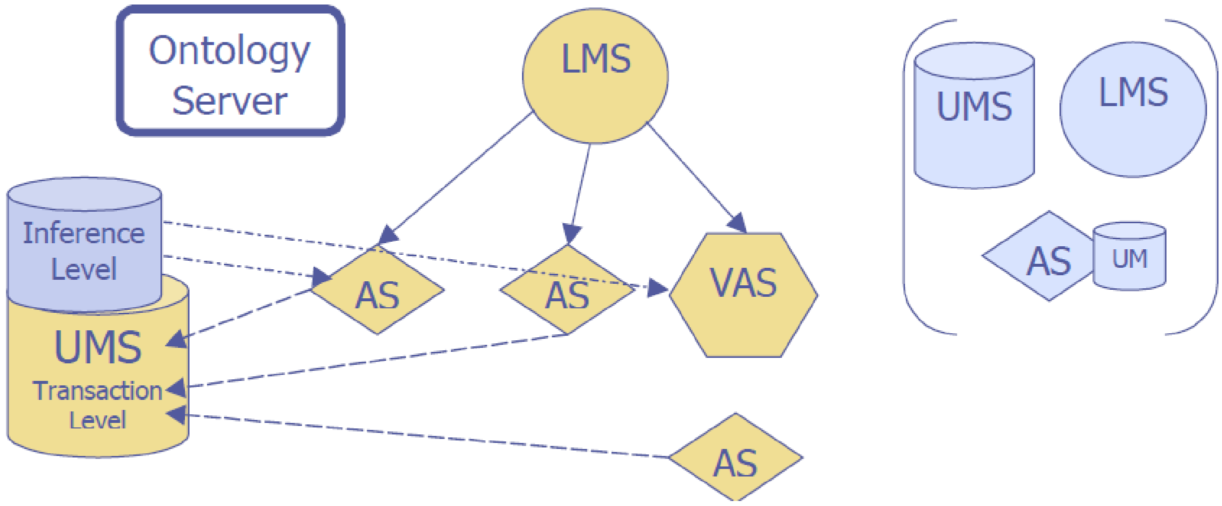


Figure 15 ADAPT2 System components by Brusilovsky et al. (2005)

The abbreviations are the following:

- LMS = Learning Management System
- AS = Activity Server
- OS = Ontology Server
- UM = User Model
- UMS = User Model Server
- VAS = Value-Added Service

The Ontology Server works as a storage for student models. Thus, the OS can consist of one ontology, but the researcher can also use any number of Ontology Servers to store users according to their ontologies. So, the student modelling is not on the side of the OS, but on the side of the UMS (User Model Server). The powerful part in this research, and in the system they developed, is that the same user (student) can easily be modelled according to different ontologies. Having any number of Ontology Servers, each containing user models and information based on their specific ontologies also provides a simple means to send or request data. Accessing or storing student data is, therefore, based on a selected ontology.

Chen & Mizoguchi (2004) researched the benefits of using an ontology to create a students' learning path. They used both a students' learning path ontology and a student model agent, and it is possible for such agents to communicate with other agents. To develop the students' learning path ontology, there were four groups of questions prepared for the agent:

1. Static information
2. Interaction data

3. Inferred information
4. Model information

As a result, they obtained fluent communication between the student model agent and other agents.

Panagiotopoulos et al. (2012) worked on developing an efficient Intelligent Tutoring System (ITS) that was used for learning from distance. They have included two taxonomies for the ontology: one part was the student's personal information and the second part was the student's academic information. The concepts that were important to the research are: learning style, modelling approach, basic characteristics of the students. Based on these, the students' learning path ontology was developed.

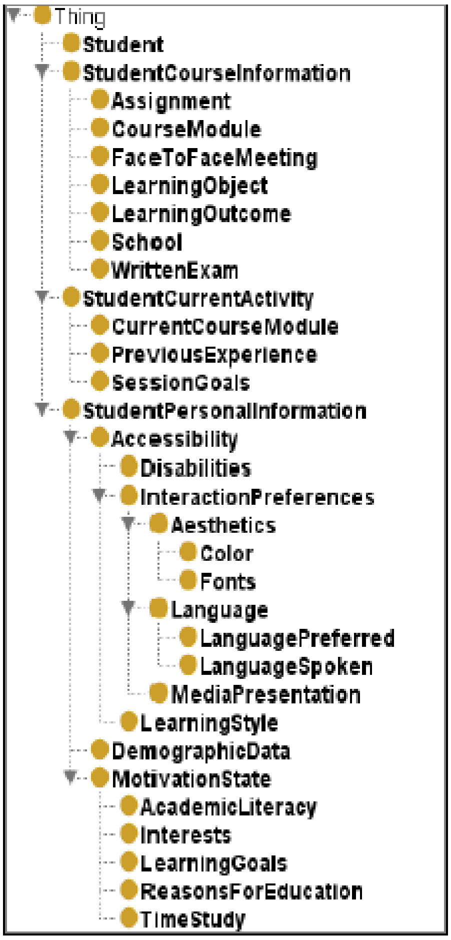


Figure 16 Ontology by Panagiotopoulos et al. (2012)

As shown in Figure 16, the ontology is based on the concepts the research used for building a students' learning path. To implement the above ontology, the researcher must define all the classes shown in the diagram. The resulting work can be a standalone ITS to be used for students that learn from distance.

4.5.3 Motivation

To be able to determine the specific weaknesses of a student and to offer him/her useful recommendations, there is a need to go deeper in understanding the relation between courses and their content as well as between each separate course. In this way, it is possible to discover the real causes of low student grades, which might be one of the pre-required courses, or a topic in a previous course (i.e. one of the Learning Objectives [LO] of the current course or old courses). The proposed solution will use the concept of ontology in representing the courses and their relations to each other or between courses and their learning objectives.

By representing the relation between courses in an ontology, the researcher can navigate the ontology when needed and know the "weak points", i.e. concepts which might be courses or learning objectives in the ontology graph.

4.5.4 Discussion

Designing the E-learning ontology

E-learning ontologies can be understood and modelled through the use of graphs where:

- There is a concept of nodes (topics, arguments, etc.) as part of an educational field of interest domain.
- The edges are binary relations that represent relations between two concepts

The focus of this design is on four kinds of relation:

- "Has Part" Relation (HP), which is a relation of inclusion
- "Has Resource" Relation (HR) a special case for the last relation, it connects between a concept and a resource explaining this concept.
- "Is Required By" Relation (IRB), which is a relation of order
- "Suggested Order" Relation (SO), which is a "weak" relation of order

As noted, two kinds of ordering relation are added to be more flexible with some fuzzy cases: IRB for strong ordering and SO for weak ordering.

For instance, if the researcher were to illustrate the development of an E-learning ontology, and this work would start by modelling a "D" educational domain:

1. First, the researcher must conceptualise the knowledge underpinning D, as well as search for a term set that represents concepts relevant to D. The previous step will result in a term list such as T = C, C1, C2, C3, where T is considered to be a potential conceptualisation of D. As for E-learning, it was acknowledged that in learning concept C, a learner would also have to learn C1, C2 as well as C3, regardless of a particular order.
2. The researcher interprets the link between a learning object and a particular concept, such as LO1 and C1, to be a HasResource relation (or HR). This relation of HasResource(C1; LO1) would mean that the meaning and content within LO1 would explain the concept in C1. As such, if the researcher would assume that C1 is the learning objective, then our corresponding E-learning experience would only consist of Learning Object 1. On the other hand, if C is our learning objective and consists of [C1, C2], then the E-learning experience could consist of multiple permutations of Learning Object 1, 2 or 3, in which the learning objective of C1 is [LO1] and the learning objective of C2 is [LO2; LO3] (i.e., the union)
3. Then the IsRequiredBy relation is interpretted, for example IsRequiredBy(C1; C2), to mean that C1 has to precede C2 in the order of learning. In this situation, if the researcher has a learning objective C, then learners need to learn through an ordered concept list [C1;C2], and they will be able to join an E-learning experience created by a Learning Object ordered sequence [LO1; LO2; LO3]. Any other alternatives such as permutations like [C2;C1] would be considered invalid.

Prediction and Advising Algorithm

The work in this component could be divided into three main steps:

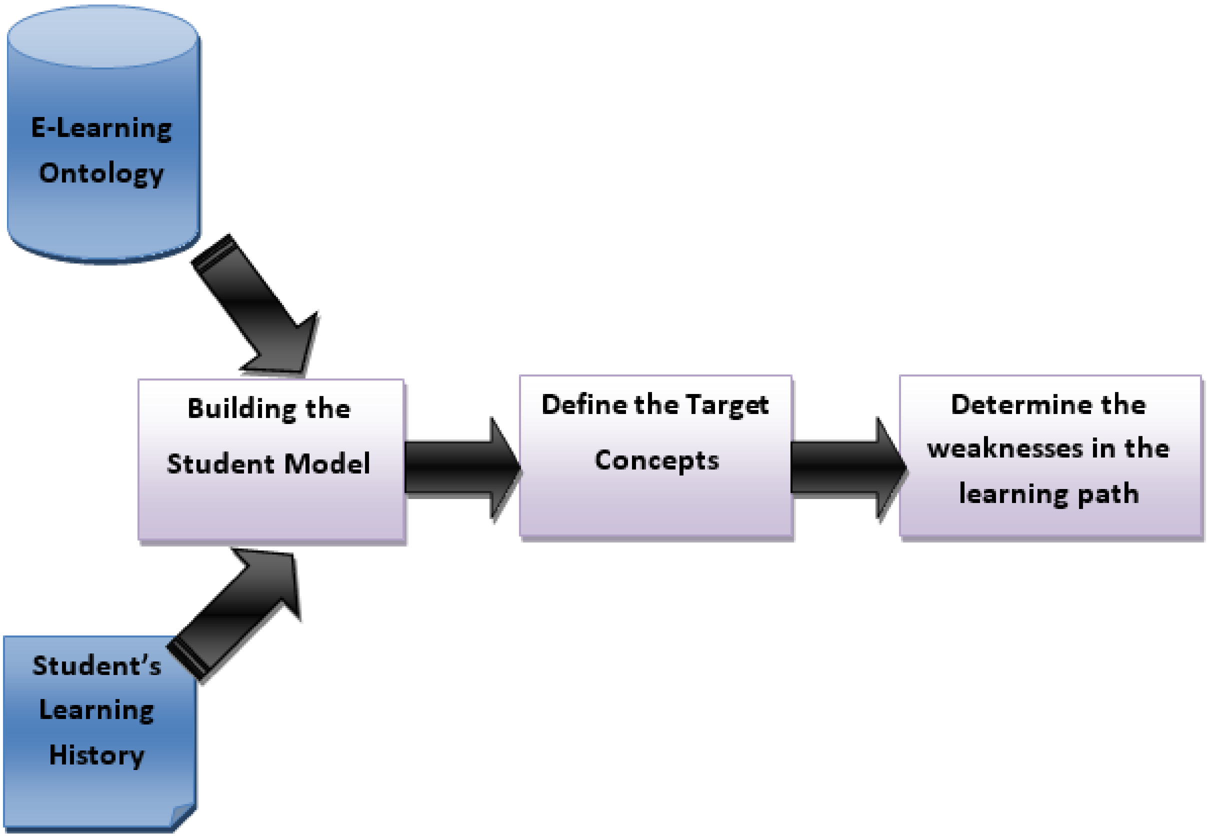

E-Learning Ontology
Building the Student Model
Define the Target Concepts
Determine the weaknesses in the learning path
Student's Learning History

Figure 17 Building the students' learning path

Step 1: Building the students' learning path

In addition to the information which has been used in previous stages, the students' learning path includes the knowledge achieved by a student and consists of a concept list of elements that each have an associated score (made of values between 0 and 1). If such a score for a particular concept would have a value greater than a particular threshold, then that score would be adequate and the concept would be considered understood.

To give a score for each concept in this model, the process starts with the learning objective. Depending on that, the score of other concepts in the model is calculated using the semantic relations (which have been described in the previous section).

For each kind of relation, a heuristic function is defined to calculate the score:

- "Has Resource" Relation (HR) and "Has Part" Relation (HP). Two possible heuristics are used: the average or the minimum (of the lower concepts in the hierarchy). An example: if X Has Part Y and Z so the score of X = Min(the score for Y, the score for Z).

- “Is Required By” Relation (IRB). Here a multiplication-based heuristic can be used which can put more importance on this relation. For instance if Y requires X, then Y would have a score equal to Y’s old score * X’s score.
- “Suggested Order” Relation (SO) which is a “weak” relation of order. Here a weighted average can be used to give a limited effect for this relation: i.e., if X is a suggested order for Y then the score of Y = the old score for Y*0.7 + the score of X*0.3. (0.3 and 0.7 are heuristic numbers)

The case of applying the minimum value is only applied on the “Has Part” relation where the parent node score is less than the score for child nodes. This option guarantees that all child concepts have a higher score when they complete their course.

Step 2: Define the Target Concepts

This is determined as follows: A set that consists of Target Concepts (TC) can be understood as a high-level concept set which needs to be forwarded and transmitted to a learner (e.g. the Learning Objective from the previous subsection) (Mangione, et al., 2010, November)

Step 3: Determine the Weaknesses in the Learning Path

The graph is navigated, which is constructed using the related concepts (a subset of concepts which are related to Target Concepts) and the relations between them. Using the scores of the concepts (the output of Step 1) the weaknesses can be determined.

4.5.5 Example

Fig. 18 shows example of the ontology adopted in this research.

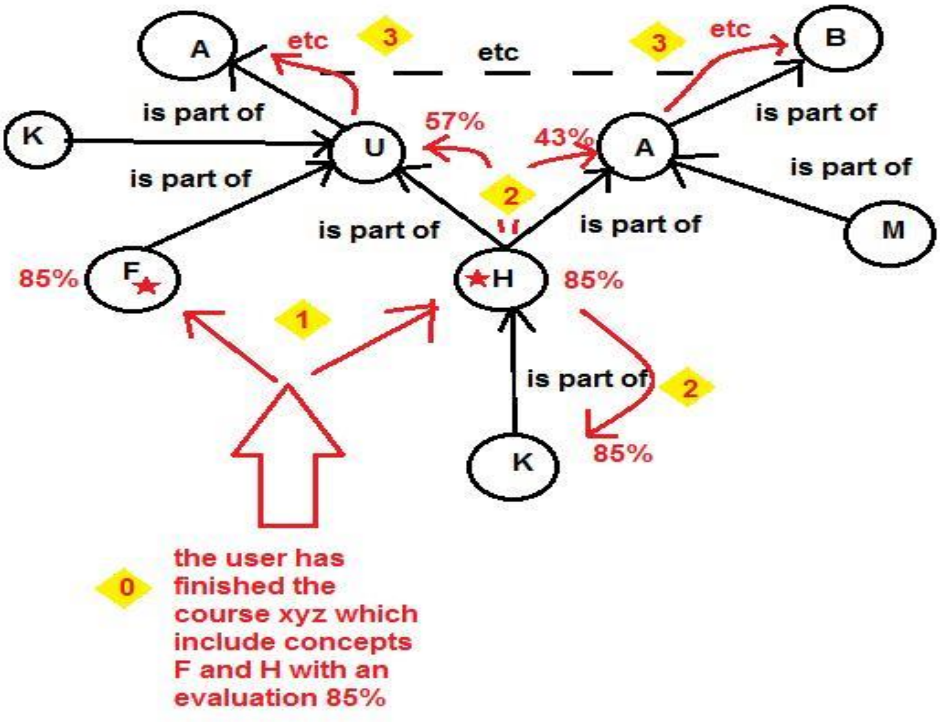


Figure 18 Ontology Example

The process is as follows:

Step 0

When the student finishes course xyz the related concepts for this course are obtained (i.e. the covered concepts in this course); in this case there are two concepts F and H.

Step 1

These concepts, F and H, would be affected by the evaluation of the course xyz. Let us say that the evaluation of the student in the course is 85%, then the evaluation of concepts F and H are also 85%.

Step 2

The change in the value of the understanding of F and H would be their parent concepts and child concepts:

K would be 85% since it is part of H

A would be 43% ((85 from H)/the number of children of A i.e. 2)

U would be 57% ((85 from H + 85 from F)/the number of children of U i.e. 3)

Step 3

The same changes would be transferred through the ontology with damping (it became smaller since we divide it by the number of the children)

4.6 Pros and Cons

This section mainly deals with the advantages and disadvantages of the proposed solution, the SPS system. While the advantages should be focused on, it is also important to evaluate the potential disadvantages as there is no perfect system.

The first part deals with the advantages in detail, and the second part discusses what the disadvantages are.

4.6.1 Pros

The proposed solution has an obvious advantage: using three approaches to build the system. As presented in this thesis, the researcher has used ID3, Hidden Markov Models and an ontology. It is definite that no approach is perfect nor very strong on its own. The three selected approaches, if they were used individually, could not have covered all aspects of certain results.

Another pro is that an advising system as this one must observe multiple sides of why the same student has a low performance. Only by evaluating all aspects can the system can come up with the best advice for the students facing more challenges or the students who had low grades and would fail in the next semester.

ID3 induces decision trees which are a summary of the knowledge contained in the data. Based on the quality of the decision tree, ID3 will reach more or less accurate conclusions. As presented in the chapter concerning our approach, the ID3-based component uses data such as the usage of students (statistics about the number of solved quizzes, number of submitted assignment, etc) in one course. As discussed in 4.5.5, SPS predicts the performance of the student by tracing the created decision tree prediction model from the root until reaching a leaf. The predicted performance of the student is the value of the reached leaf. Furthermore, the SPS system demonstrates to the student the causes of this predicted performance by detecting the attributes that lead to this route through the decision tree prediction model. It would also display to the student the values that they should have in these detected attributes to get higher performance. The students will therefore know the reason behind the predicted performance. Researchers will also need to consider the personality of the user to be more adaptive. In other words, it important to give advice which considers the student's learning style. Many cognitive studies show that each person has a different learning style. Cognitive research shows that when learners learn with their preferred learning style, they will take less time to master a given concept in a more comprehensive way. So, determining the learning style, helps the system in finding the most suitable advice for each student. The user's personality is part of the reason why the researcher included the Hidden Markov Models. This research classify students based on their preferences. There are basically three categories: Visual, Auditory and Kinesthetic. The researcher has also used a questionnaire built from 30 questions with three possible answers per question. As presented, this research has a set of rules to classify the students and to train the Markov model.

Then the researcher has noticed that including both ID3 and Hidden Markov Models, there are also other causes for low grades in the case of many students. To cover the other side of the cause, the researcher included the ontology for observing learning paths. The researcher analysed the relations between courses and the content of the courses, and also the relations between the courses and their learning objectives. Using the relations for the built ontology, the weakness will be determined and better advice can be given to students. Adding the ontology on top of the other two components results in a more robust system, able to assess performance in a more efficient way, as presented in this thesis.

The three different components of the system use three different algorithms. At the beginning of this chapter, the researcher has presented the works of other scholars. It is important to see that they have used these algorithms in their research. To make sure that they created something useful, the researcher has reviewed both their results and whether their works were cited by others or not. It was found that the works were cited and also effective. This was shown by the tables, graphs and other visual representations that were presented in each of the papers. It is obvious that every researcher must have included a demonstration that their solution works. The demonstration is where it possible to see mathematical details to reveal how performance and accuracy have been improved. Using a new approach based on existing approaches that have already been proved can only lead to further success. While each of the scholars revealed how some parts remain uncovered or how some parts of the algorithms might have weaknesses, they did obtain promising results.

Another pro is the interaction between the proposed system and the student. As soon as enough student data is collected the performance prediction is in action. This means that a message is generated, so that both the student and the teacher will know how to take action next.

In the next chapter the researcher will evaluate our system to measure the performance and the results. It is important to make sure it is a system that brings impact and a positive change. In other words, It is important to measure how useful the proposed solution is. The researcher will also use statistics and visual representations to make sure it is possible demonstrate that the solution works in a useful manner.

4.6.2 Cons

The researcher has not studied much about evaluating motivation. As research referenced by this thesis also indicates, lack of motivation can easily cause low grades. Identifying the weakness of lack of motivation would not yet be enough to assess solutions. A system can be built which can, for example, evaluate more sides of each students' personalities and suggest certain activities that would partially engage the students. Also, teachers would need to counsel certain students, giving them opportunity to gain more motivation. Such counselling is not easy, and must be adaptive because what causes one student to be receptive can easily cause another student to keep lacking motivation.

It would also be interesting to further check which aspects of ID3, HMM and Ontology could be improved in our case. Using several approaches to each and measuring results by statistics could open up new doors and new horizons to potentially new solutions. It is also not guaranteed that another approach would not improve the system the researcher already designed. As an idea, Bayesian networks can have potential as they have been used successfully by researchers time after time. As resulting from previous works about Bayesian Networks, it is a matter of future work to evaluate how a Bayesian Network could possibly improve on our current results as a fourth component or as a complementary part of one of the existing components. More scholars concluded that using Bayesian Networks helped them achieve more accuracy in student performance prediction.

Another negative side is that the researcher was unable to question students and teachers likewise, asking them what they would change or improve in our system. While some of their answers would be irrelevant to improvement, other answers would be able to open up new methods of implementing new parts of the system or changing some parts of it completely.

Yet another disadvantage is that this research has not considered that the system could be used by students with special needs. While other students might understand our system in a certain way, it is not guaranteed that the special needs students will obtain the same results. The reason is that their personality is not the same, and their needs are not the same. Thus, it is highly probable that this system, left unchanged, will not necessarily improve or accurately predict the performance of students with special needs. For future work, Researchers should assess the differences between such systems and systems specially designed for special needs students. Understanding related work and how other scholars have approached the issue will help us improve this system in the future and create two versions of it for the two categories of student.

It is also a disadvantage that this research has not considered extras to be included for excellent students. While excellent students maintain their results, it might prove useful to provide them new ways through this software to broaden their horizons and study even more.

4.7 Summary

Predicting students' performance using software is a challenge. To understand all aspects of such work, it was first important to comprehend the problem itself and the various approaches used by researchers in past years. As the general problem in the area revealed, it was always difficult to make sure that all criteria contribute to the final results. While several articles successfully covered an approach on one or two sides of the problem, they have all failed to include the other aspects. One of the greatest challenges in decisions was to decide for or against certain algorithms used in the area. The researcher has presented three main algorithms and the ways these algorithms can be used to approach the problem.

The chapter first discusses the general problem with popular approaches. The researcher then present ID3 which can be used successfully to predict the students' level of understanding. This research has also reviewed other approaches that have implemented ID3-based algorithms. It is obvious that ID3 decision trees contribute to the final solution, as these are effective in solving one side of the problem. C4.5 is as an extension of ID3 but was not used here. The main reason for omitting C4.5 is that the developed system uses multiple algorithms to predict student performance. Any aspects that were not necessarily covered by ID3 are covered by the other components. Furthermore, as Joseph & Devadas found in 2015, scholars can modify ID3 in ways that it will have more accuracy than C4.5

Hidden Markov Models mostly deal with a student's learning style. There are three very important types of learner: Visual, Auditory and Kinesthetic. As the researcher has studied the literature and have reviewed in the usage of the developed system, a student will never have 100% of one learning style. If it is noted Visual as V, Auditory as A and Kinesthetic as K, then V+A+K = 100%. In other words, a student may be 65% auditory, but that means they are also 15% visual and 20% kinesthetic, or any other possible amount that satisfies the condition above.

The reason for including HMMs in the study is because ID3 will never be able to evaluate a learning style. Yes, the researcher could have formulated certain quiz strategies and ask strategic questions, but there is never a guarantee that ID3 would eliminate the necessity of HMMs. Furthermore, the hidden states cannot be covered by decision trees. The ID3 algorithm is sequential, in the sense that it uses a series of decisions based on certain values to reach conclusions. Such predefined decisions, even if cleverly stated, are not enough to cover all the aspects of the Hidden Markov Model. In the results of other scholars who have implemented HMMs, the researcher can see that they were able to improve previous results. Learning about their findings and how HMMs helped in obtaining higher accuracy, it is fair to consider HMMs as a component in the SPS system, used to evaluate the different learning styles.

HMMs also have a useful property related to one of the goals: creating a system that can be adaptive. A clever modelling will always lead to open doors for adaptive approaches. While scholars covered performance prediction in many different ways, using multiple algorithms and strategies, there have been mentions of the fact that the future work should focus on an adaptive system. The intention of having the SPS is also to make that happen, bringing forward an adaptive solution that would give advice based on how each student had their results, and also based on their other characteristics such as their learning styles.

The learning style was determined using 30 questions. In this chapter the researcher has included the questions with their possible answers included so that one can understand how the style was determined. After the questions are answered, researches need to represent the styles. The researcher has included details about how the classification rules work. In the end of the classification, researches also need to take V, A and K into account. It is always enough to measure by V, A and K because every learning style is a sum of these three main styles. Creating sub-groups of Visual, Auditory or Kinesthetic is out of scope, because even if we had sub-groups, the conclusions and final results would lead to the three main groups. The last part of how a HMM works is how the model is trained. The researcher presented details as to how the Markov model was trained.

The third important component that our system focused on is using the proper ontology. Ontology as a concept is very simple, as it is only about representing relationships. Building useful students' learning paths using an ontology is not necessarily simple.

One of the basic problems when modelling students by ontology is what to consider and what to ignore. It is easy to make mistakes in terms of including or excluding information, thus it was important to review previous works related to this area. The researcher has reviewed several ontology-related works (Day et al., 2005, July; Paneva, 2006; Pramitasari, et al., 2009). The interesting part of the review was to find that not only Hidden Markov Models, but also ontologies were used to create an adaptive students' learning path. If the researcher has ID3 already, which is a strong decision making system based on decision trees, and add the power of two adaptive components such as HMM and ontologies, researchers can potentially get closer to the adaptive system students and tutors need.

Ontology is a strong method to evaluate the personality of students. In current work, the researcher has evaluated relations between courses and their content, and also between courses and their learning objectives.

The SPS system uses all three components, based on ID3, HMM and Ontology. It was merely a matter of personal decision to not include Bayesian Networks. As reviewed in what the scholars have studied, Bayesian Networks were successfully used in multiple works. This research focused on a personalised system that did not require a fourth component.

In order to confirm that a system makes sense and can be used efficiently, the first question is whether it’s necessary for students and tutors or not. The following are the main reasons why the researcher believe that the system is necessary:

1. Other scholars developed similar systems. If their papers have been published and cited, used by different people, it is already proof that their work was useful.
2. In every work the researcher has reviewed and used as a reference before building the system, clear graphical and statistical data is shown to prove that the respective works have a higher accuracy and performance than the preceding works. It is important to know if something is improved or not, because a system is only useful if it is better

than an older system. It is never necessary to build a new wheel that is just as efficient as the best wheel already available on the market.

3. Upon evaluating the SPS system by contacting the UOD (University of Dammam), the researcher learned that the student group that used our system improved their results at the end of the second semester. Comparing the same group of students to another group of students that had no access to our system, it was easy to measure that there was significant improvement. For more details see chapter 5, section 5.2.2.

The researcher did not prepare any questionnaire for the students, but has measured the obtained results. Whether the students would acknowledge it or not, the SPS prediction system gave valuable advice to most of them, except the excellent students. As it is presented in this work, excellent students tend to maintain their excellence, thus a decline in their performance is not an issue.

Tutors in universities always face challenges when it comes to the semester-end results of their students. While some tutors have enough experience to counsel their students, other tutors do not have enough experience. The following are some of the main reasons why tutors would benefit from the use of the SPS system in their universities:

1. Students study more or less based on various social and non-social factors. While excellent students always find their way around, other student groups tend to decline as this research evaluation also revealed. If students have a steady performance or improvement in their performance, the tutors can also feel a relief in their efforts. For a tutor it is important to know that their students have learned something, and have more knowledge after each semester passes.
2. Universities always struggle to keep a good reputation. Generally, the reputation of a given university is not only the name of the university or the professionalism of their tutors, but also the results of the students that call the university their alma mater. Thus, an effective performance prediction system helps tutors counsel their students and understand their personalities as well. Many times counselling is not possible without the knowledge of the personality, and the SPS gives further insight into that.

Using systems such as SPS is something new for the students. Burnout is often caused because students feel bored or disengaged. Using our system provides something new, and when tutors present the system to their students, stronger connections will be made and students will have a boost in engagement because they will feel more competitive. When there is a public measurement system, it is human nature to try and become better than all the others. Tutors can become more successful if their students are successful

Chapter 5 Evaluation

5.1 Introduction

This chapter covers the stages that were followed to evaluate the system. It also covers types of data collected and used by the software. The chapter begins with a section that discusses data collection and then moves on to how the researcher tested the software on a group of 100 students in UOD. Finally, the chapter ends with a summary of the obtained results.

5.2 Experimental Results and Observations

5.2.1 Collecting data

The data used by SPS was collected from the E-learning system of the University of Dammam (UOD). This university was established in 1395H / 1975G as a national university aiming at spreading higher education in the eastern area of Saudi Arabia. Distance learning in the UOD has only recently been initiated and is still under development. The Deanship of Distance Learning was established on 11/5/1425H. The reason for the selection of the E-learning system of UOD is that it is the only available system for the researcher. The researcher asked for data from another university, KFU, but the request was refused.

In the UOD, the E-learning systems used are the E-learning Management Electronic System (EMES) and the Virtual Classroom System (CENTRA). EMES is an integrated computer system that manages the educational process where this system aims to facilitate the process of interaction between student and faculty member. Its features are course development, ease of use, Arabic support, ability to assess students, communication between student and faculty, quality of scientific content design and using the latest technology for educational means, developing self-learning among students, easy management and support of the educational process. CENTRA provides lectures on the Internet based on the smart classroom environment and essential elements needed by both the instructor and the student. CENTRA is one of the widely adopted Distance Learning systems. It provides the required modules for a dynamic interaction between the student and the lecturer. It also provides open interaction during the learning process.

CENTRA provides the following functions:

- The ability to start Live Virtual Classrooms
- Real time interaction between students and instructors
- Audio and visual learning content
- Allows students to share applications
- Provides functions for testing and assessment
- Provides multimedia tools – Teleconferencing, Video, Voice Over IP
- Supports many languages
- Supports live e-Meetings

The data used in this thesis originates from the EMES database. EMES stores students' profiles, instructors' profiles, course information, data about the students' visits, the assignments data, and the quizzes data in its database. In detail, the data stored in EMES for each course is as follows:

1. Students' names
2. Students' IDs
3. Students' genders
4. Instructors' genders
5. Instructors' names
6. Instructors' grades
7. Courses' classes
8. Courses' names
9. Number of students in each class
10. Total time spent in the E-learning system
11. Total time spent in the E-learning system in seconds
12. Number of visits to the system
13. Number of read pages
14. Number of visits to content
15. Number of submitted assignments
16. Number of downloaded assignments
17. Total grade of the assignments
18. Number of solved quizzes
19. Number of correct answers

20. Number of wrong answers
21. Total grade of the quizzes

From all these data, student names, student IDs, course names, data about the students visits, the assignments data, and the quizzes data are selected as attributes for the data mining process to predict students' performance. These data are chosen because they give the best prediction models as noted from the comparison in chapter 2. So, the final data used is the following:

1. Students' IDs
2. Total time spent in the E-learning system
3. Number of visits to the system
4. Number of read pages
5. Number of submitted assignments
6. Number of total assignments
7. Number of average assignments
8. Number of solved quizzes
9. Number of total quizzes
10. Number of average quizzes
11. Performance

Student usage of more than one course was obtained, namely: Algorithm Analysis, Introduction to Algorithms, Data Structures, Object Oriented Programming, Advanced Algorithms and Introduction to Artificial Intelligence. The Algorithm Analysis course is selected as a case study because it is the most intense course; it has more students' usage data stored in the EMES database compared to the other courses. There are 100 students registered in the course, divided into two classes. The usage of these 100 students is selected.

It is worth noticing that this concept is related to Algorithms and Data Structures, Computer Programming, Mathematical Concepts, Mathematical Problem Solving, and Applied Mathematics. Also, there is an ID and Parent ID for each concept. In this way, all concepts in this file are connected.

5.2.2 Experimentation

It is important to evaluate the software to find if it will achieve accurate results in predicting students' performance. As a result, the researcher considered two methods to evaluate the system:

a) Compare the software with other software that predicts students' performance
b) Test the software on a number of students

It was decided that the best approach is to test the software on a number of students in the UOD. Evaluation has been performed using the 3rd-year students in the computer science faculty at UOD. Two groups of students were chosen: Group A and Group B, each of which had 50 students. In the evaluation, the researcher selected a set of related subjects: Algorithm Analysis, Introduction to Algorithms and Data Structures, Object Oriented Programming, Advanced Algorithms, and Introduction to Artificial Intelligence.

The year is divided into two terms with each term carrying a weight of 50%. 60% of the total marks are allocated to written examinations at the end of each term and the remaining marks are allocated for continuous assessments during the term. The grading system is based on total marks obtained as shown in the table below:

Table 7 The grading system of total marks

Marks	**Grading**
90-100	Excellent
75-89	Very Good
60-74	Good
50-59	Fair
< 50	Fail

For evaluating the impact of the system, the researcher noted the performance of both groups in the first term without any usage of the system. In term two the researcher gave the students of group A advice with the help of our system, and kept group B without any advice. The performance of both groups was noted for both the terms and results, and are shown in the next table:

Table 8 The performance of both group 1&2 in 1st term

Examination Results at End of 1st Term				
Marks	**Grading**	**Group A**	**Group B**	**Total**
90-100	Excellent	4	3	7
75-89	Very Good	11	14	25
60-74	Good	18	15	33
50-59	Fair	15	15	30
< 50	Fail	2	3	5
Total		50	50	100
Examination Results at End of 2nd Term				
Marks	**Grading**	**Group A**	**Group B**	**Total**
90-100	Excellent	4	4	8
75-89	Very Good	15	11	26
60-74	Good	22	17	39
50-59	Fair	8	16	24
< 50	Fail	1	2	30
Total		50	50	100

The figure below shows the Performance Matrix for students belonging to group A and B over the two term period.

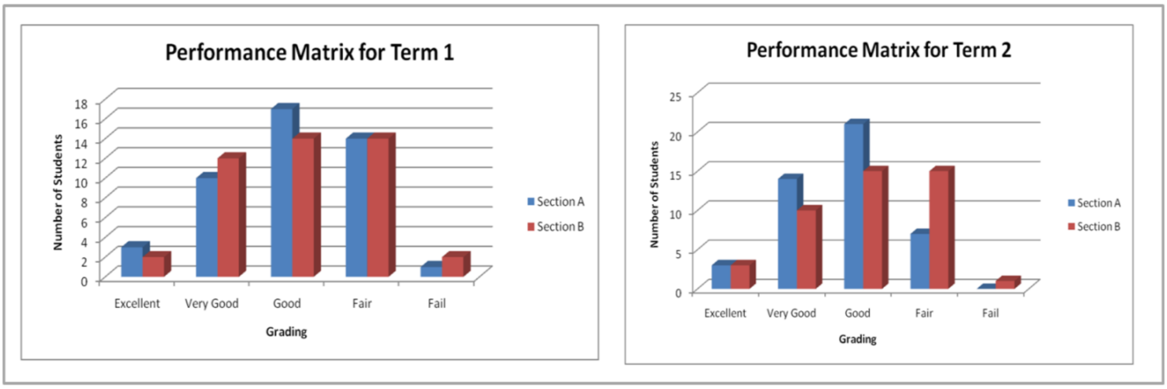


Figure 19 Performance matrix for both terms

As can be seen from this figure, there is a definite improvement in the performance of students in Group A who have been using the system throughout the next term. To further analyse the impact of the learning system on the students, the average performance and standard deviations are derived from the two sets of students. In relation to the above mentioned improvement, one cannot ignore the Hawthorne Effect, which means that the people who know they are part of a study may act differently as a consequence of their consciousness and awareness. This may mean that the observed groups of students tried to study harder, being fully conscious of the fact that they are under observation and part of a research project. Thus, the Hawthorne Effect may prove a limitation of the current research: it is unknown what the results would have been if participants had not known that they were under observation.

To evaluate the standard deviation, first, the researcher took the mean value of each range mentioned in the last table above and indicated it by the symbol 'X' in the next two tables. So, for example, the mean for the range of marks from 90 to 100 is 95. Variables 'Y1' and 'Y2' respectively denote the frequency distribution, i.e. the number of students in that range for Group A and Group B respectively. The standard deviation represents the variability of scores about the average in each group. Higher average scores with lower variability will be considered as a positive outcome of the system. It is positive because it will not only result in higher average scores but also will show that students getting lower marks have moved into a closer band about the mean score.

After one whole year, the average scores of students from Group A performance increased significantly by +5% circa from 67% to 72% while Group B had negligible increase from 67.82% to 68.12% circa .03%, confirming that it favoured the average students. As it is expected, the performance of bright students remained consistent over the two terms for both the groups.

Table 9 The average scores of students

	Before	After	The change %
Group A	67.15	72.02	+5%
Group B	67.82	68.12	+03%

The observation of the changes in the grades of the students from both groups was possibly due to collecting student grades before, during and after the usage of the software. It is clear that the greatest changes in terms of improvement occur for the Good and Very Good students (these categories are observable above in this section, in the presented tables). The Excellent students tend to keep their standards with or without intervention, while Good and Very Good students tend to improve more if there is intervention. By intervention one should understand the usage of the proposed system. Thus, by seeing the performance prediction through the software, the students seemed to be more motivated towards improving their results – and that is what occurred for Group A students, as shown in this section before.

The actions taken together with the students was first to give them access to the proposed system and then to allow them and their tutors to use the suggestions (performance prediction and mentioned suggestions as presented in this thesis) for their own benefit. This means that the students were presented with the resulting options, they were informed about the features and possibilities and then they were able to use it for one year. The same was with the tutors – they had the explanations and presentations, and were able to monitor progress and see how students are doing with the help of the proposed system.

5.3 Analysis of Evaluation

5.3.1 Introduction

This section deals with the analysis of the evaluation. The quality of the evaluation should be assessed in order to understand why it is efficient in the case of this system.

5.3.2 The analysis

The information gathered from the UOD was successfully used to both train and test the SPS system which have presented and developed. It was important to test the method with real-life data, and not with data that was randomly created in favour of our own method.

Thanks to including student information from UOD, it is possible to see how the system would evaluate a situation that is very similar to any other university. If there was more contribution, the researcher could have tested the system for other universities too.

The students from UOD have agreed to test our software for a semester. Based on the statistical data obtained from two groups the researcher has classified in the previous section, it is can clearly be seen the effectiveness of the SPS system. As it was expected, it is not the excellent students who need such systems. Their level of study is consistent, and with or without using the SPS, their results would still be the same.

Thankfully, it was possible to see the impact on the group, especially for the average students. It is important to observe the decline in the results of the students from the group that didn't use the SPS system. Except the excellent students from Group B – which didn't have access to SPS – all students experienced a change in their performance in the second semester. Based on the results of the Group A students – who were given access to SPS – the researcher know that Group B would possibly have also improved results if they only had access to our system.

It is also very important to notice that the researcher has sorted the data that was obtained from the EMES database. Based on the previous presentations in this chapter, it is known that not all data is useful for evaluation.

Using the standard deviation was important to understand which areas improved. As already concluded, the average students benefit the most. Using the SPS, it was both statistically and visually observable that Group A performed better than Group B, thanks to the presented solution.

5.3.3 Hawthorne Effect

As stated in terms of potential improvement, this study did not assess the potential flaws or changed results due to the Hawthorne Effect. In other words, the subjects may have acted differently being in full knowledge of the fact that they were involved in an experiment.

The Hawthorne Effect is discussed by scholars from different approaches, such as (McCarney, et al., 2007; Wickström & Bendix, 2000). Thus, one cannot fully conclude that the improvement was at a certain level, as the Hawthorne Effect was not considered. As part of future research, the scholar proposes an improvement to measure the potential flaws brought by this, and find the ways to measure improvement correctly. Also, as found by later research, there needs to be more specific assessment. Different situations should optimally define a specific effect. However, such definitions and psychological study is out of the scope of this thesis.

5.4 Summary

As revealed by both our evaluation and common sense and understanding of how universities are in 2016, the researcher can state that the SPS advising system is an improvement, and while it is not completely new, it is new in terms of performance and accuracy. Thus, gradually, more and more universities can benefit from its usage. Personalised advice giving is valuable for the academic sector, and the work provides exactly that.

In the evaluation the researcher has used the data from the EMES database. The researcher had to filter the data, making sure to use what is relevant. The evaluation had two important parts, from which one was gathering the data and the other part was measuring the results. The researcher gathered the data for the purpose of training the system. The evaluation was conducted on 3rd year students from the UOD and this work used Group A and Group B for two semesters. These are two student groups, from where only Group A had access to the SPS system and Group B was left without the aid of the system that have been developed.

In the evaluation, the grading system used is based on total marks obtained as shown in the table below:

Table 10 The grading system of total marks

Marks	Grading
90-100	Excellent
75-89	Very Good
60-74	Good
50-59	Fair
< 50	Fail

It is could easily be observed and measured that the SPS system helped Group A. While all excellent students from both groups have maintained their excellent results, Group B suffered in their study results from semester one to semester two. Thus, the students who didn't have the excellent results could benefit from SPS. There was a significant improvement in Group A, where students were given advice based on what the SPS predicted.

It should be noted that the researcher ensured that Group A did not receive more attention than Group B and the only variable that changed was the use of SPS. This ensured that the change in the students' performance was a result of using SPS and no other variable contributed in raising their performance.

Using the standard deviation, it was clear that the most benefit comes for the average students, namely those students with average grades. Lastly, due to not considering the Hawthorne Effect, the real improvement may be different from the measured values.

Chapter 6: Conclusions and Future Work

6.1 Introduction

This chapter concludes the work conducted throughout this thesis, and provides details about future work related to this area of research. It is known that any contemporary work can be improved in the following years, but it is of great significance to comprehend the ways and means by which enhancement is truly made possible.

Section 6.1 presents the outcome of the research conducted, and section 6.2 demonstrates the obstacles. Section 6.3 covers the strengths and contribution to the literature. Finally, section 6.4 points out several research directions in the future, while section 6.5 summarises this chapter.

6.2 Summary of the thesis

This thesis addresses the issues surrounding the use of data mining within E-learning systems. It introduces the basic concepts of data mining and how it can be used in improving such systems. It provides a data mining process for predicting students' performance based on archived data extracted from an E-learning system. By identifying the future prediction of the students' performance early enough, the student will have the chance and motivation for improvement. On the other hand, it could utilise the predictive information and analysis reports to catch up with both short term actions to adapt the pedagogy for what is left in the semester and a longer term adaptation of the course design to suit those categories of students.

A model of student performance prediction is proposed. This model employs the decision tree classification data mining technique on data archived in LCMS systems about student usage and performance of the previous course intakes. An algorithm is also designed to trace the predictive decision tree model in order to generate advisory improvement strategies of study for each student. In addition, a prototype (SPS) was developed in Java, and using Weka, to evaluate the proposed model of prediction and advising. Although the collected data was not large, it was sufficient to assess the viability of the SPS model. The prediction approach used in this thesis is easily adaptable to different types of courses, different population sizes, and allows for different attributes to be analysed.

Hidden Markov Models (HMM) were one of the significant components of the proposed prediction system. The most significant motivation of using HMMs was the potential to give advice which considers a student’s learning style. To make it possible, a method was necessary to review the learning styles. For this purpose, the presented questionnaire was used. The result of using the questionnaire is the possibility to add more features which represent the learning style.

Secondly, as another addition to the developed HMM usage was classifying based on VAK properties. Briefly, the system was developed so that it can classify students in categories T1, T2, T3, T4, T5, T6 and T7. The hidden states of the presented model use the V, A, and K components. The transition matrices were defined and a description was given as to how transitions from one state to another are conducted. Lastly, to assure the model is working, the HMM was trained. Through generic rules and examples it is clear how the training process leads to a powerful and adaptive component within the system.

Ontologies were used to build the third component of the proposed prediction system. A description of the used ontology rules presents all necessary details. The ontology built relies on four relations as presented in the System Design chapter: Has Part (HP), Has Resource (HR), Is Required By (IRB) and Suggested Order (SO).

Three main steps assure the correct setup of the E-learning ontology: building the students' learning path, defining the target concepts and determining the weaknesses in the learning path. The resulting ontology, using the described relations, allowing both strong and weak ordering relations, provides a students' learning path that includes more details than merely the ID3 and HMM components.

As noted and demonstrated in the previous chapters, the work was tested in real life situations. The students who were allowed to use the prediction system versus the students who had no access proved a significant positive impact in their study results. Thus the outcome of the conducted research is not only the three-fold design based on the three components, but also evidence that it improves grades of students who use the proposed solution. Furthermore, the three-fold component design managed to fill in gaps in previous works related separately to ID3 decision tree models, HMM models and ontologies. Using these techniques separately makes it impossible to fill in certain gaps. Briefly stated, it is clearly an improvement and innovative in the area of data mining in eLearning. While there has been research in data mining for eLearning, there are no extensive studies in the area and the existing studies did not take applications to levels as high as the current thesis. In the Literature Review chapter it becomes clear how related works in the literature had their gaps, lack or insufficient application that left room for improvement.

There are a few lessons learned throughout this research, which can be summarised as follows:

- Data mining can be used in E-learning environments and can enhance them.
- Students' performance can be predicted using data mining on archived data of previous semesters.
- Students' performance prediction should help students to improve their performance.
- Students' performance prediction helps the instructors to find ways to adapt the course structure for better learning processes for those target categories of students.

6.3 Obstacles

Although the research has achieved its aim and objectives, there are limitations in the research methodology process and findings. However, only one set of data was obtained, namely end-of-semester performance data only. Accordingly, monthly prediction models were impossible to produce, and hence, only one prediction model was generated. Additionally, because the E-learning program at DU is still in its early stages, the obtained data set was only for a small number of students (100 students). As a result, the prototype had to be modified to utilise this set of data for both model building and model verification. In addition, one of the main limitations of the study is that the system has not been tested on large cases. Also there is a need to prepare the E-learning ontology which consists of the courses and the learning objectives, etc, which is believed to need more effort to be done effectively.

The essence of decision trees and how they are constructed brings limitations in design. Essentially, decision trees are limited because they do not provide adaptive behaviour. Once a tree is designed, it can only advance and make decisions based on its ramifications, which are also pre-designed. As discussed in earlier chapters, scholars did propose and describe modified decision tree-based algorithms. An interesting alternative design would be considering the replacement of the ID3 component by a Bayesian Network component. To make that work with the other components is out of the scope of presenting current obstacles, as it would change many aspects of the core design, and the results. Whether it would bring an improvement or not cannot be made clear at this point. Research must be conducted to state whether a well-built Bayesian component is a better replacement for the ID3 component or not. Bayesian Networks can be trained, or in other words they can gain new knowledge. Gaining new knowledge opens a door to new ideas, and to multiple potential ways of transmitting the knowledge to the Bayesian Networks that could be a valuable component of this system. Also, various scholars have used BNs (Bayesian Networks) in their work, of which part was directly related with predicting student performance, and other works were in different areas. While ID3 is partially fixed, meaning it does not learn anything on the way, BNs are flexible, because they are "teachable", which we call training.

Lacking a questionnaire that would ask both students and teachers is an obstacle. Before building a system, a creative idea could have been surveying people, and asking specific questions, such as:

- Do you find student performance prediction necessary? Motivate your answer
- If you ever used a student performance predictor, what do you think it should provide as functions? Are all the mentioned functions necessary?
- Would you enjoy an automatic system that advises you to take actions towards your academic improvement? What sort of advice do you find helpful?

The list of questions is not complete, and it is out of the scope of the current discussion. However, it is clear that the mentioned survey questions could have brought new ideas in designing the system. Obviously, this does not conclude that the three-fold component design would not be sufficient, but it does suggest that improvements would have been made possible by knowing what people think.

6.4 Strengths and Contribution to the Study of Knowledge

The findings highlighted before have made an innovative contribution to the practical knowledge in the field of eLearning and data mining. The study contributes in filling the gap in existing literature relevant to using data mining in E-learning. Using ML and data mining techniques allowed to give more accurate prediction and advice to both students and instructors.

The primary strength of the design is a consequence of using ID3, HMM and ontologies. While these three approaches have disadvantages and gaps if separated and studied, blending them together in the way presented in this thesis fills many of the gaps and turns disadvantages into advantages. For example, while ID3 would not be able to provide adaptability, HMM and ontologies step into adapting behaviour based on students' personalities and learner styles. Reviewing the related literature, it is clear that no scholar used such a three component design as the one proposed in this thesis. Thus, as future work, the improvements revealed by the current work will help scholars to gain new knowledge and apply that knowledge more effectively.

As reviewed in the previous chapters, there have been numerous learner style models proposed by scholars. It is known that some models prevailed and are in contemporary use. One of the disabilities in any of the models is that it only covers style based on certain points of view or based on certain preliminary assumptions. Although the assumptions are documented and taken further into research, there has been research that attempted to reveal how certain models are very limited or too simplistic to ever describe a student's actual personality and learning style. The scholar's work is contribution to the lacking description, because the result of the three presented components uses a student ontology that covers multiple aspects of personality, of how much V, A or K a student has, and ultimately of a more complex learner style.

Furthermore, limited research has been conducted by scholars to determine the relationship between tutor styles and student styles, and also to determine how the tutor styles would be able to influence positively or negatively how students learn. While the previous research suggested already that there is an important relationship and also teachers must adapt their styles, there has not been enough conclusion and application. The current thesis fills in the gap of advisory systems targeted bilaterally: towards students and teachers likewise. Certainly, previous research states the possibility, but neither of them provides a system as complex and proven helpful as the proposed system design. By addressing both parties, the chances to improve academic performance enhance due to being informed and gaining more motivation based on a system that merely predicts the performance of the students.

6.5 Future Work

The outcomes of the study results and the conclusions attempt to suggest a number of recommendations: There are several promising issues to extend the work presented in this thesis:

1. Applying the model online.
2. Predict student performance periodically (weekly or at least monthly).
3. Apply the model using large dataset.
4. Give instructors advice to help them in improving course structure.

Each area of the outlined future work will now be described. In each part more detail, suggestions and discussion will clarify how and why the mentioned improvements should take place in near future.

6.5.1 Applying the Model Online

It is not necessary to repeat the description of the developed model here, neither of its overall functions and possibilities. For such details, refer to the previous chapters that give a detailed insight. Here, it will be considered why applying the developed model online should be taken seriously, and what aspects or challenges this future work would bring.

The emerging of the online world in countries worldwide is a fact. Thus, knowing the facts about the expansion of the World Wide Web, and how more people rely on online resources, it is clear that completing something online is an important consideration.

In academic settings, the use of technology, which includes computers and usage of eLearning, an online prediction system would be useful. To make sure the model would function online, it is necessary to assess:

1. Security
2. Responsive design
3. Uptime (should be accessible at least 99.99% of times)
4. Expandability
5. Robustness

6. User-friendliness
7. Reliability
8. Role(s) of administrators
9. Implementation of the model's algorithms to function online
10. Timely feedback and timely results

The above list may be incomplete, but the most important aspects are covered. The scholar will now discuss the specifics related to the above list, where necessary. Topics as uptime or responsive design are considered simple, not requiring further clarifications.

Security

In the online platforms, security threats prevail and there is a global tendency towards cyber criminality. It is possible to achieve high levels of security. To enhance security beyond limits, one could attempt the following:

- Pay people money if they find vulnerabilities.
- Hire somebody to attempt all types of known attacks, starting with cookie manipulations and up to SQL injections or session attacks.
- Be informed about the contemporary website attack methods and even about spyware, malware, Trojans and viruses that could affect either the server hosting the modules or the website itself (without spreading to the entire server).

User-friendliness

As well as software, online platforms must be easy to use, and overall user-friendliness is essential. The friendlier it is for users, the more they will choose to use such a product. This property is also given by assuring the following:

- Simple words on the user interface
- Suggestive graphics
- Menus to be arranged in logical order, using submenus where appropriate
- Easily interpretable error messages, prediction results, informational messages

- Help/How to use built using simple English terms, so that the non-technical students would also understand

Also, for students the user-friendliness is one of the main factors of influence, whether a platform is online or offline.

Reliability

Reliability would be assured by daily usage of thousands of students, over a period of a few years. If the users can confirm that both quality and efficacy have been observable while using, then one can conclude that the online platform is reliable. Reliability, obviously, relates to many of the aspects described above.

Role(s) of Administrators

Administrators have more control in an offline or online platform than ‘regular’ users. They can add, remove or modify user accounts, control databases and various other tasks. Students also need to know there is an administrator to turn to when necessary.

Implementation of the Model’s Algorithms to Function Online

The current implementation is a matter of languages and architectures, but then several other factors gain importance, such as: speed, server load, and response time. When implementing, certain elements might need change or brand new design to allow functioning optimally. The fact that some people accessing the online platform have outdated computers that run slower and on older operating systems is noteworthy and part of the online design to assure quality.

Timely Feedback and Timely Results

An implementation can be excellent, yet not provide timely feedback or timely results. Algorithm complexity and run times must be thoroughly tested, and big server loads must be simulated. More research and design is needed to assure that even the mentioned factors are solved.

6.5.2 Predict Student Performance Periodically

Student performance prediction, as discussed, is important in education both for the tutors and the learners. The future work question is: how often is it necessary to predict performance? What factors determine the selected time period?

In other scholars' work, it has been discussed that the prediction is done for year-end or semester-end results. There has not yet been a detailed discussion on whether it is necessary or not to predict monthly or weekly.

While it would be convenient to say monthly without discussion, it is not enough for improvements. Universities in every course have certain weekly assignments. Some courses bring up only bi-weekly assignments, but it is impossible for a student to have merely bi-weekly frequency of homework or different projects. Given the fact that students are provoked to prepare themselves every week by studying the subjects and applying theory by solving homework or by working on group projects, a weekly indication of performance would be good for most real life applications.

To further expand on the topic, a weekly performance prediction report allows students and tutors to assess progress statistically on a monthly basis. Related strictly to this section of proposed future work, the scholar should also examine what statistical insights the weekly reporting brings and how each of these insights can be used towards other improvements of the model and the system as well. Knowing weekly progress, one will already know monthly progress easily. This allows anyone to handle special cases in time, because after a few weeks it is already visible if some students face difficulties in their studies.

6.5.3 Apply the Model Using Large Dataset

Kokkelenberg et al. (Kokkelenberg, 2008) discussed what effects would the size of a class have on student performance. They have used a large dataset to test their model.

Strauss and Volkwein (Strauss, 2002) have also used large datasets in their study for testing. They measured the differences in student performance between students who attended courses for two years and those who attended courses for four years.

Thus, the literature is not yet sophisticated in the area of applying a student performance prediction model using large dataset.

The motivation of applying the model using large datasets raises many aspects. The very first aspect deals with measuring the accuracy and usefulness of the built system in case of access to larger datasets. While it was clearly stated in this thesis that the used dataset was enough to prove the validity of the model, it is even more powerful to assess the same on a larger scale. In the process of testing and observing, as part of this application, new ideas may easily come to light. Some of the potential ideas have to do with:

- Student performance in universities across multiple countries, including poor and excellently developed regions
- Student performance based on the specialisations and courses within a university
- How statistics from one academic setting may help another similar academic setting

Furthermore, large datasets allow more observations that have to do with how students and tutors find this solution helpful or not. By allowing them to fill questionnaires with very specific questions asked, the scholar gains more knowledge with the user feedback, allowing for more research and improvements on all levels. User feedback will obviously be more accurate when the model is applied on larger datasets.

6.5.4 Give Instructors Advice to Help Them in Improving Course Structure

In the current state of the proposed system, it is only possible to give advice strictly related to the students' performance prediction. In other words, the model could quickly be expanded to give advice regarding low-performing students and average students, or students that went from good to average or from average to low-performing. While such details would be incorporated without difficulties, there is no necessary correlation between the measured performance and the course structure.

To relate courses structure to the proposed work, future work should analyse what else is needed and how it is incorporated in order to allow the system to advise each instructor personally. At least one way to assess performance based on the structure of a course should be included in the future research.

Advising can be done by collecting information. The simplest is collecting the outcome of a certain course, such as how well each student performed at the end of the semester. This simple statistic already tells part of the problem. Another implementation would be collecting anonymous student feedback related to a course or even to an instructor. Such feedback can be gathered stating that is for statistics or to assess the success of each course. Thus, students will feel empowered to answer, by knowing they lose nothing – they don't have to write their names and they can affirm things very generally, but to the point. Then, it is simple to collect all this data and use it to advise the instructors personally. An automatic part can only be achieved – meaning an automatic advising sent to instructors – by measuring grades, and then sending out a generic message based on a mathematical decision. For example, if the average grade is above X but below Y, the message is M1, if it's above Y and below Z, the message is M2 and so forth.

Furthermore, an important question to ask is whether or not courses structure is the only factor to deal with student performance. As mentioned in this thesis, tutor styles also impact learner achievements. Thus, the future work must measure tutor styles too and make sure that advice is given based on multiple aspects, of which course structure is just one. Finding more aspects to it is another required improvement.

6.5.5 Comparison with existing systems

SPS has shown that it can possibly predict student performance effectively. Testing it on a group of students showed that their performance has changed and their marks increased. However, there is a need to compare the developed approach with other software that predicts student performance to ensure that the developments are competitive with the state-of-the-art. In addition, this will allow SPS to improve through the evaluation of the strength and weakness of other systems.

To compare, ideally, would mean to select several alternative software solutions. When those are selected, the only way to truly compare is to select different student groups for each software solution, and monitor the improvement in each group. Preferably, each group should have similar ages and similar study profiles.

One of the notable works that compares such software is (Hsieh & Cho, 2011). In the mentioned paper, a framework has been developed to determine whether one system is more efficient than the other. For the study, there were 783 students from seven higher education institutes in Hong Kong.

As the comparison was based on four hypotheses, it resulted that three out of the four were supported by the results. ISI means instructor-student interactive learning tools, and SP means self-paced learning tools. The four hypotheses were:

1. "Information quality of ISI e-learning tools is higher than that of SP e-learning tools."
2. "The perceived usefulness of ISI e-learning tools is higher than that of SP e-learning tools."
3. "Learning outcomes of ISI e-learning tools are higher than those of SP e-learning tools."

4. “Learner satisfaction with instructor-student interactive e-learning tools will be higher than learner satisfaction with self-paced e-learning tools.”

The research concluded that the ISI approach will outperform the SP approach in perceived usefulness, learning outcomes, student satisfaction and information quality.

6.6 Research limitations

As is true for any system, such systems as the one presented here will always leave room for improvement. Because of the time given to conduct this study and the nature of the sample used in this study there are some limitations in this research. Hence, it would be interesting for other researchers to find detailed answers to the following questions, which would help to identify new paths to take in the future:

1. What do students think regarding the utility of SPS?
2. What would students say that SPS lacks or could include as a new function?
3. Do students find new motivation upon using our system?
4. If a university decides to implement our system and use it every year, what are their statistics? Are their students constantly improving performance? How much of an improvement do they detect?
5. What would tutors suggest based on their experience of using our system? Does it have all necessary functions? Does it need something new? Does it have something that we should remove?

The above list of questions could be continued, but even with the above lists the answers can help in future work.

6.7 Summary

Throughout the current thesis a novel approach to predicting student performance has been presented, measured and put in practice. The three-fold component design suggested in the thesis that was capable of intertwining the strengths of three different approaches is very powerful. While it was clear that the scholar brought genuine ideas and additional knowledge to the fields of data mining, eLearning, student modelling, and predictive educational systems, it was also challenging to see how this work opens up new doors to future research in areas that do not yet have the same amount of academic interest or solid conclusions leading to the same principles.

Certain strengths were obvious in the current design, which also led to contribution to knowledge in this area. Due to various obstacles, certain aspects could not be discussed or analysed in enough detail. Regardless of the obstacles that were presented, the research was shown to be effective and useful, through the experiments and statistics presented in the previous chapters.

The final goal to build a system which is able to predict student performance in a novel way has been achieved. The current design, as any other, can further be improved. However, it is clearly an improvement based on the literature review. It is promising and door-opening to see that the proposed system design is helpful in academic contexts, measured in real life on real students.

As a conclusion, through many challenges and hard work, it was possible to bring new knowledge to several areas of research. Future work and research can be based on the findings of the current thesis, as it is a novel and detailed study, discussing both strengths and weaknesses of the used methods.

References

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting Student's Performance Using ID3 and C4.5 Classification Algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 3, No.5, September 2013.

Aggarwal, C. C., & Reddy, C. K. (2013). Data clustering: algorithms and applications. CRC Press.

Aldhafeeri, F., & Male, T. (2015). Investigating the learning challenges presented by digital technologies to the College of Education in Kuwait University. Education and Information Technologies, 1-11.

Alghamdi, F. and Jensen, R., (2014). Latest Trends in Data Mining for E-learning Systems. International Journal of Information Technology & Computer Science, 17 (2), 49-60

Amershi, S., & Conati, C. (2009). Combining Unsupervised and Supervised Classification to Build User Models for Exploratory. JEDM-Journal of Educational Data Mining, 1(1), 18-71.

Amershi, S., Conati, C., & McLaren, H. (2006). Using feature selection and unsupervised clustering to identify affective expressions in educational games. Workshop in Motivational and Affective Issues in ITS, 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan.

Amershi, S., Conati, C., & McLaren, H. (2006). Using feature selection and unsupervised clustering to identify affective expressions in educational games. Workshop in Motivational and Affective Issues in ITS, 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan.

Anaya, A. R., & Boticario, J. G. (2009). A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks. International Working Group on Educational Data Mining.

Anaya, A. R., & Boticario, J. G. (2009). A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks. International Working Group on Educational Data Mining.

Arellano, J.B., Divine, A.S., Dobes, Z.K. and Liu, G., SBC Technology Resources, Inc., 2004. System and methods for an architectural framework for design of an adaptive, personalized, interactive content delivery system. U.S. Patent 6,694,482.

Baldi, P. and Brunak, S. (1998). Bioinformatics. Cambridge, Mass.: MIT Press.

Baradwaj, B. K., & Pal, S. (2012). Mining Educational Data to Analyze Students' Performance. arXiv preprint arXiv:1201.3417.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, 37(6), 1554-1563.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, 37(6), 1554-1563.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 41(1), 164-171.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 41(1), 164-171.

Beal, C. R., Mitra, S., & Cohen, P. R. (2006). Modeling student engagement with a tutoring system using Hidden Markov Models.

Beal, C. R., Mitra, S., & Cohen, P. R. (2006). Modeling student engagement with a tutoring system using Hidden Markov Models.

Beal, C., Mitra, S., & Cohen, P. (2007). Modeling learning patterns of students with a tutoring system using Hidden Markov Models. Artificial Intelligence in Education, R. Luckin et al. (2007), IOS Press.

Beal, C., Mitra, S., & Cohen, P. (2007). Modeling learning patterns of students with a tutoring system using Hidden Markov Models. Artificial Intelligence in Education, R. Luckin et al. (2007), IOS Press.

Benson, L., Elliot, D., Grant, M., Holschuh, D., Kim, B., & Kim, H. (2002). Usability and instructional design heuristics for e-Learning evaluation. P., & S. (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2002 (pp. 1615-1621).

Bhardwaj, B., & Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.

Bhat, S. (2004). Generalization of ID3 algorithm to higher dimensions.

Billings, K., & Moursund, D. (1988). Computers in education: An historical perspective. ACM Sigcue Outlook, 20(1), 13-24.

Billings, K., & Moursund, D. (1988). Computers in education: An historical perspective. ACM Sigcue Outlook, 20(1), 13-24.

Birch, C. J., & Clements, M. (2003). Engaging Small to Medium Sized Enterprises in Learning.

Blunsom, P. (2004). Hidden markov models. Lecture notes, August, 15, 18-19.

British Educational Research Association (BERA) (2011), Ethical Guideline for Educational Research, Revised Ethical Guidelines. Exeter. Author. Available: https://www.bera.ac.uk/wp-content/uploads/2014/02/BERA-Ethical-Guidelines-2011.pdf?noredirect=1.

Brookfield, S. D. (2009). Self-directed learning. . International handbook of education for the changing world of work (pp. 2615-2627). Springer Netherlands.

Bruffee, K. A. (1999). Collaborative learning: Higher education, interdependence, and the authority of knowledge. .

Bruno, N. C. (2001). STHoles: a multidimensional workload-aware histogram. ACM SIGMOD Record (Vol. 30, No. 2, pp. 211-222). ACM.

Bruno, N. C. (2001). STHoles: a multidimensional workload-aware histogram. ACM SIGMOD Record (Vol. 30, No. 2, pp. 211-222). ACM.

Brusilovsky, P., Sosnovsky, S., & Yudelson, M. (2005). Ontology-based framework for user model interoperability in distributed learning environments. World Conference on E-learning in Corporate, Government, Healthcare, and Higher Education (Vol. 2005, No. 1, pp. 2851-2855).

Bu, F., Hao, Y., & Zhu, X. (2011, July). Semantic relationship discovery with wikipedia structure. In IJCAI Proceedings-International Joint Conference on Artificial Intelligence (Vol. 22, No. 3, p. 1770).

Bullen, M. (2014). What is eLearning? (http://dspace.col.org/bitstream/handle/11599/665/eLearning-Transcript.pdf?sequence=3&isAllowed=y).

Bush, T. (2007) ‘Authenticity in research–reliability, validity and triangulation’, in Briggs, A. R. J., Coleman, M. and Morrison, M. (ed.) Research methods in educational leadership and management, Sage Knowledge. DOI: http://dx.doi.org/10.4135/9781473957695.n6.

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. Evolution of teaching and learning paradigms in intelligent environment (pp. 183-221). Springer Berlin Heidelberg.

Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to E-learning problems. Evolution of teaching and learning paradigms in intelligent environment (pp. 183-221). Springer Berlin Heidelberg.

Cen Li and Jungsoon Yoo. Modeling Student Online Learning Using Clustering. In ACM-SE 44: Proceedings of the 44th Annual Southeast Regional Conference, pages 186–191. ACM, 2006.

Chaudhuri, S., Ganti, V., & Kaushik, R. (2006). A primitive operator for similarity joins in data cleaning. 22nd International Conference on Data Engineering (ICDE'06) (pp. 5-5). IEEE.

Chaudhuri, S., Ganti, V., & Kaushik, R. (2006). A primitive operator for similarity joins in data cleaning. Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on (pp. 5-5). IEEE.

Chen, M., Chen, Y. and Liu,Y. (2007) Learning Performance Assessment Approach Using Web-Based Learning Portfolios for E-learning Systems, IEEE, In Proceedings of the Fifth

IEEE International Conference on Advanced Learning Technologies, Vol. 37, No. 6, PP. 557 – 561

Chen, W., & Mizoguchi, R. (2004). Learner model ontology and learner model agent. Cognitive Support for Learning-Imagining the Unknown, 189-200.

Clark, R. (2002). Six principles of effective e-Learning: What works and why. The e-Learning Developer's Journal, 1-10.

Coffield FJ, Moseley DV, Hall E and Ecclestone K (2004). Should we be using learning styles? What research has to say to practice. London: Learning and Skills Research Centre/University of Newcastle upon Tyne.

Cohen, L., Manion, L., and Morrison, K. (2013) Research methods in education, 7th revised edition, London: Routledge.

Conner, R.F., 1980. Ethical issues in the use of control groups. New Directions for Program Evaluation, 1980(7), pp.63-75.

Corbett, A. C. (2005). Experiential learning within the process of opportunity identification and exploitation. . Entrepreneurship Theory and Practice, 29(4), 473-491.

Courville, K. (2011). Technology and Its Use in Education sector: Present Roles and Future Prospects. Online Submission.

Creswell, J. W. (2013) Research design: Qualitative, quantitative, and mixed methods approaches. USA: Sage publications, Inc.

D’Antoni, S. (2009). Open educational resources: Reviewing initiatives and issues.

Das, D., Singh, N. K., & Sinha, A. K. (2006). A comparison of Fourier transform and wavelet transform methods for detection and classification of faults on transmission lines. Power India Conference, 2006 IEEE (pp. 7-pp). IEEE.

Dash, M. &. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.

Dash, M. &. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.

Day, M. Y., Lu, C. H., Yang, J. T., Chiou, G. F., Ong, C. S., & Hsu, W. L. (2005, July). Designing an ontology-based intelligent tutoring agent with instant messaging. Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on (pp. 318-320). IEEE.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 1(2), 1542-1552.

Downs, E., Gruse, G.G., Hurtado, M.M., Lehman, C.T., Milsted, K.L. and Lotspiech, J.B., International Business Machines Corporation, 2001.Electronic content delivery system. U.S. Patent 6,226,618

Dunn, R. S., & Dunn, K. J. (1978). Teaching students through their individual learning styles: A practical approach. Prentice Hall.

Durlauf, S. N., Kourtellos, A., & Minkin, A. (2001). The local Solow growth model. European Economic Review, 45(4), 928-940.

Durlauf, S. N., Kourtellos, A., & Minkin, A. (2001). The local Solow growth model. European Economic Review, 45(4), 928-940.

Ellis, R. (2004). Down with boring E-learning! Interview with E-learning guru Dr. Michael W. Allen. Learning circuits. Retrieved from www.astd.org/LC/2004/0704_allen.htm

European working session on learning, E. (2010). Machine learning. (Journal, magazine, 1900s) [WorldCat.org]. [online] Worldcat.org. Available at: http://www.worldcat.org/title/machinE-learning/oclc/39741651?referer=di&ht=edition [Accessed 28 May 2016].

Felder, R. M. (2005). Understanding student differences. . Journal of engineering education, 94(1), 57-72.

Flick, U. (2015). Introducing research methodology: A beginner's guide to doing a research project. Sage.

Frymier, A. B., & Houser, M. L. (2000). The teacher- student relationship as an interpersonal relationship. . Communication Education, 49(3), 207-219.

Fukunaga, K. &. (1970). Application of the Karhunen-Loeve expansion to feature selection and ordering. IEEE Transactions on Computers, 19(4), 311-318.

Fukunaga, K. &. (1970). Application of the Karhunen-Loeve expansion to feature selection and ordering. IEEE Transactions on Computers, 19(4), 311-318.

Gagnes, R. (1997).The Conditions of Learning and Theory of Instruction, New York: Holt, Rhinehart and Winston.

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Switzerland: Springer.

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Switzerland: Springer.

Gilakjani, A. P. (2011). Visual, auditory, kinaesthetic learning styles and their impacts on English language teaching. Journal of Studies in Education, 2(1), 104-113.

Gilakjani, A. P. (2011). Visual, auditory, kinaesthetic learning styles and their impacts on English language teaching. Journal of Studies in Education, 2(1), 104-113.

Gilakjani, A. P. (2012). A match or mismatch between learning styles of the learners and teaching styles of the teachers. International Journal of Modern Education and Computer Science, 4(11), 51.

Gilakjani, A. P. (2012). A match or mismatch between learning styles of the learners and teaching styles of the teachers. International Journal of Modern Education and Computer Science, 4(11), 51.

Gilbert, S. M., & Jones, M. G. (2001). E-Learning Is E-Normous Training over the internet has become the fastest-growing workplace performance improvement tool-and utilities are using it in several ways. Electric Perspectives, 66-85.

Giunta, A. A. (2002). Use of data sampling, surrogate models, and numerical optimization in engineering design. AIAA paper, 538, 2002.

Gonzalez, H., Han, J., & Shen, X. (2007). Cost-conscious cleaning of massive RFID data sets. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 1268-1272). IEEE.

Hamdy, H., Williams, R., Tekian, A., Benjamin, S., El Shazali, H., & Bandaranayake, R. (2001). Application of" VITALS": visual indicators of teaching and learning success in reporting student evaluations of clinical teachers. EDUCATION FOR HEALTH-ABINGDON-CARFAX PUBLISHING LIMITED-, 14(2), 267-276.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd edition). Waltham, MA: Morgan Kaufmann Publishers (Elsevier).

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd edition). Waltham, MA: Morgan Kaufmann Publishers (Elsevier).

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT Press.

Hase, S., & Kenyon, C. (2000). From andragogy to heutagogy. . Ultibase Articles, 5(3), 1-10.

Hawk, T. F., & Shah, A. J. (2007). Using learning style instruments to enhance student learning. Decision Sciences Journal of Innovative Education, 5(1), 1-19.

Hawk, T. F., & Shah, A. J. (2007). Using learning style instruments to enhance student learning. . Decision Sciences Journal of Innovative Education, 5(1), 1-19.

Hawk, T. F., & Shah, A. J. (2007). Using learning style instruments to enhance student learning. Decision Sciences Journal of Innovative Education, 5(1), 1-19.

Hellerstein, J. M. (2008). Quantitative data cleaning for large databases. United Nations Economic Commission for Europe (UNECE).

Hofstede, G. (2010) Cultures and Organizations: Software of the Mind, 3rd edition, USA: McGraw Hill Professional.

Holden, M. T., & Lynch, P. (2004). Choosing the appropriate methodology: understanding research philosophy. The marketing review, 4(4), 397-409.

Honey, P., & Mumford, A. (1992). The manual of learning styles. Peter Honey Publications; 3rd edition (January 31, 1992).

Hsieh, P. A. J., & Cho, V. (2011). Comparing e-Learning tools' success: The case of instructor–student interactive vs. self-paced tools. Computers & Education, 57(3), 2025-2038.

Hu, Q., Yu, D., & Xie, Z. (2006). Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern recognition letters, 27(5), 414-423.

Iman, R. L. (1999). Appendix A: Latin Hypercube Sampling. Encyclopedia of Statistical Sciences, Update Volume 3, Wiley, NY, 408-411.

Itmazi, J. A., & Megías, M. G. (2005). Survey: Comparison and evaluation studies of learning content management systems. Unpublished manuscript.

Itmazi, J. A., & Megías, M. G. (2005). Survey: Comparison and evaluation studies of learning content management systems. Unpublished manuscript.

Jensen, R., & Shen, Q. (2007). Fuzzy-rough sets assisted attribute selection. Fuzzy Systems, IEEE Transactions on, 15(1), 73-89.

Jensen, R., & Shen, Q. (2007). Fuzzy-rough sets assisted attribute selection. Fuzzy Systems, IEEE Transactions on, 15(1), 73-89.

Jeong, H., & Biswas, G. (2008, June). Mining Student Behavior Models in Learning-by-Teaching Environments. EDM (pp. 127-136).

Jeong, H., & Biswas, G. (2008, June). Mining Student Behavior Models in Learning-by-Teaching Environments. EDM (pp. 127-136).

Johns, J., & Woolf, B. (2006). A Dynamic Mixture Model to Detect Student Motivation and Proficiency. In Proceedings of the National Conference on Artificial Intelligence (Vol. 21, No. 1, p. 163). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Johnstone, I. M., & Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association.

Journal of Computer Assisted Learning 26, 243-257.

Joseph, S., & Devadas, L. (2015). Student's Performance Prediction Using Weighted Modified ID3 Algorithm. International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882, Volume 4, Issue 5, May 2015.

Kay, J., Halin, Z., Ottomann, T., & Razak, Z. (1997). Learner know thyself: Student models to give learner control and responsibility. Proceedings of International Conference on Computers in Education (pp. 17-24).

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. The Internet and higher education, 8(1), 13-24.

Knowles, M. S. (1970). The modern practice of adult education (Vol. 41). . New York: New York Association Press.

Kokkelenberg, E. C. (2008). The effects of class size on student grades at a public university. Economics of Education Review, 27(2), 221-233.

Krause, U. M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on E-learning in statistics. Learning and Instruction,19(2), 158-170.

Kuhn, T. L. (2008). Historical foundations of academic advising.

Kuljis, J., & Liu, F. (2005). A Comparison of Learning Style Theories on the Suitability for elearning. Web Technologies, Applications, and Services, 2005, 191-197.

Larose, D. T. (2014). Discovering knowledge in data: an introduction to data mining. John Wiley & Sons.

Lemieux, C. (2009). Monte carlo and quasi-monte carlo sampling. Springer Science & Business Media.

Lemieux, C. (2009). Monte carlo and quasi-monte carlo sampling. Springer Science & Business Media.

Li, D., Wang, S., & Li, D. (2016). Spatial Data Mining: Theory and Application. Springer.

Li, D., Wang, S., & Li, D. (2016). Spatial Data Mining: Theory and Application. Springer.

Liaw, S. S. (2008). Investigating students’ perceived satisfaction, behavioural intention, and effectiveness of E-learning: A case study of the Blackboard system. Computers & Education, 51(2), 864-873.

Mangione, G. R., Gaeta, M., Orciuoli, F., & Salerno, S. (2010, November). A Semantic Metacognitive Learning Environment. 2010 AAAI Fall Symposium Series.

Matthews, D. (1999). The origins of distance education and its use in the United States. THE Journal (Technological Horizons In Education), 27(2), 54.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. BMC medical research methodology, 7(1), 1.

Merriam-Webster Dictionary. (2016). Retrieved from www.merriam-webster.com/dictionary/ontology

Muller, C. (2001). The role of caring in the teacher-student relationship for at-risk students. . Sociological inquiry, 71(2), 241-255.

Murphy, K. P. (2002). Dynamic bayesian networks. Probabilistic Graphical Models, M. Jordan, 7.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling, 9(4), 599-620.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling, 9(4), 599-620.

Muthukrishnan, S., & Strauss, M. J. (2013). Washington, DC: U.S. Patent and Trademark Office. Patent No. U.S. Patent No. 8,600,704.

Najafabadi, M. O., Poorsadegh, M., & Mirdamadi, S. M. (2013). Challenges of Application ICTs in Technical and Vocational Training from Students’ and Instructors’ Perception in Maragheh. International Journal of Advanced Science and Technology, 54, 105-111.

Najafabadi, M. O., Poorsadegh, M., & Mirdamadi, S. M. (2013). Challenges of Application ICTs in Technical and Vocational Training from Students’ and Instructors’ Perception in Maragheh. International Journal of Advanced Science and Technology, 54, 105-111.

Neuhauser, C. (2002). Learning style and effectiveness of online and face-to-face instruction. The American Journal of Distance Education.

Neuhauser, C. (2002). Learning style and effectiveness of online and face-to-face instruction. The American Journal of Distance Education.

Neuhauser, C. (2010). Learning style and effectiveness of online and face-to-face instruction. The American Journal of Distance Education.

Nichols, M. (2003). A theory of eLearning. Educational Technology & Society, 6(2), 1-10.

Noor, K. B. M. (2008). Case study: A strategic research methodology. American journal of applied sciences, 5(11), 1602-1604.

Ogunde, A., & Ajibade, D. (2014). A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm. Journal of Computer Science and Information Technology March 2014, Vol. 2, No. 1, pp 21-46.

Oluwatayo, J. A. (2012). Validity and reliability issues in educational research. Journal of Educational and Social Research, 2(2), 391-400.

Ontology_(information_science). (2016). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Ontology_(information_science)

Onwuegbuzie, A. J. (2000) 'Expanding the Framework of Internal and External Validity in Quantitative Research', Paper presented at the Annual Meeting of the Association for the Advancement of Educational Research (AAER), Ponte Vedra, FL, USA.

Oswald, C., Ghosh, A. I., & Sivaselvan, B. (2015). An Efficient Text Compression Algorithm-Data Mining Perspective. Mining Intelligence and Knowledge Exploration (pp. 563-575).

Paechter, M., Maier, B., & Macher, D. (2010). Students' expectations of, and experiences in E-learning: Their relation to learning achievements and course satisfaction. Computers & education, 54(1), 222-229.

Panagiotopoulos, I., Kalou, A., Pierrakeas, C., & Kameas, A. (2012). An ontology-based model for student representation in intelligent tutoring systems for distance learning. Artificial Intelligence Applications and Innovations (pp. 296-305). Springer Berlin Heidelberg.

Paneva, D. (2006). Use of Ontology-based Student model in Semantic-oriented Access to the Knowledge in Digital Libraries. proc. of HUBUSKA Fourth Open Workshop "Semantic Web and Knowledge Technologies Applications", Varna, Bulgaria (pp. 31-41).

Pardos, Z., Heffernan, N., Anderson, B., & Heffernan, C. (2007). The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks. In User modeling 2007 (pp. 435-439). Springer Berlin Heidelberg.

Pashler, H., McDaniel, M., Rohrer, D. and Bjork, R., 2008. Learning styles concepts and evidence. Psychological science in the public interest, 9(3), pp.105-119.

Phoon, K. K., Huang, S. P., & Quek, S. T. (2002). Implementation of Karhunen–Loeve expansion for simulation using a wavelet-Galerkin scheme. Probabilistic Engineering Mechanics, 17(3), 293-303.

Pivec, M., & Baumann, K. (2004). The role of adaptation and personalisation in classroom-based learning and in e-learning. Journal of Universal Computer Science, 10(1), 73-89.

Pivec, M., & Baumann, K. (2004). The role of adaptation and personalisation in classroom-based learning and in E-learning. Journal of Universal Computer Science, 10(1), 73-89.

Pivec, M., Dziabenko, O., & Schinnerl, I. (2004). Game-based learning in universities and lifelong learning: “UniGame: social skills and knowledge training” game concept. Journal of Universal Computer Science, 10(1), 14-26.

Pivec, M., Dziabenko, O., & Schinnerl, I. (2004). Game-based learning in universities and lifelong learning: “UniGame: social skills and knowledge training” game concept. Journal of Universal Computer Science, 10(1), 14-26.

Popescu, E., Trigano, P., & Badica, C. (2007). Towards a unified learning style model in adaptive educational systems. Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007) (pp. 804-808). IEEE.

Popescu, E. (2010). Adaptation provisioning with respect to learning styles in a Web-based educational system: an experimental study.

Pramitasari, L., Hidayanto, A. N., Aminah, S., Krisnadhi, A. A., & Ramadhanie, A. M. (2009). Development of student model ontology for personalization in an E-learning system based on semantic web. International Conference on Advanced Computer Science and Information Systems (ICACSIS09), Indonesia, December (pp. 7-8).

Quellec, G., Lamard, M., Erginay, A., Chabouis, A., Massin, P., Cochener, B., & Cazuguel, G. (2016). Automatic detection of referral patients due to retinal pathologies through data mining. Medical Image Analysis, 29, 47-64.

Quellec, G., Lamard, M., Erginay, A., Chabouis, A., Massin, P., Cochener, B., & Cazuguel, G. (2016). Automatic detection of referral patients due to retinal pathologies through data mining. Medical Image Analysis, 29, 47-64.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1.

Ramaswami, M., & Rathinasabapathy, R. (2012). Student Performance Prediction. International Journal of Computational Intelligence and Informatics, 1(4).

Raspopovic, M., Cvetanovic, S., & Jankulovic, A. (2016). Challenges of Transitioning to E-learning System with Learning Objects Capabilities. The International Review of Research in Open and Distributed Learning, 17(1).

Rogowsky, B. A., Calhoun, B. M., & Tallal, P. (2015). Matching learning style to instructional method: Effects on comprehension. Journal of Educational Psychology, 107(1), 64.

Rogowsky, B. A., Calhoun, B. M., & Tallal, P. (2015). Matching learning style to instructional method: Effects on comprehension. Journal of Educational Psychology, 107(1), 64.

Rokach, L. and Maimon, O. (2008). Data mining with decision trees. Singapore: World Scientific.

Rossel, R. V., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158(1), 46-54.

Saunders, M. and Lewis, P. (2011) Doing research in business and management: An essential guide to planning your project, USA: Trans-Atlantic Publications, Inc.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. The Journal of educational research, 99(6), 323-338.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. The Journal of educational research, 99(6), 323-338.

Sebastiani, P., Ramoni, M., Cohen, P., Warwick, J., & Davis, J. (1999). Discovering dynamics using Bayesian clustering. International Symposium on Intelligent Data Analysis (pp. 199-209). Springer Berlin Heidelberg.

Sharp, J. G., Bowker, R., & Byrne, J. (2008). VAK or VAK- uous? Towards the trivialisation of learning and the death of scholarship. Research Papers in Education, 23(3), 293-314.

Sharp, J. G., Bowker, R., & Byrne, J. (2008). VAK or VAK- uous? Towards the trivialisation of learning and the death of scholarship. Research Papers in Education, 23(3), 293-314.

Siemens, G. (2014). Connectivism: A learning theory for the digital age.

Simpson, O. (2013). Supporting students in online open and distance learning. Routledge.

Soller, A. (2007). Adaptive support for distributed collaboration. The adaptive web (pp. 573-595). Springer Berlin Heidelberg.

Soller, A. (2007). Adaptive support for distributed collaboration. The adaptive web (pp. 573-595). Springer Berlin Heidelberg.

Soller, A., & Lesgold, A. (2007). Modeling the process of collaborative learning. The role of technology in CSCL (pp. 63-86). Springer US.

Soller, A., & Lesgold, A. (2007). Modeling the process of collaborative learning. The role of technology in CSCL (pp. 63-86). Springer US.

Sra, S., Nowozin, S. and Wright, S. (2012). Optimization for machine learning. Cambridge, Mass.: MIT Press.

Stratonovich, R. L. (1960). Conditional markov processes. In R. L. Stratonovich, Theory of Probability & Its Applications (pp. 156-178).

Stratonovich, R. L. (1960). Conditional markov processes. In R. L. Stratonovich, Theory of Probability & Its Applications (pp. 156-178).

Strauss, L. C. (2002). Comparing student performance and growth in 2-and 4-year institutions. Research in Higher Education, 43(2), 133-161.

Surjono, H.D., 2011. The design of adaptive e-Learning system based on student's learning styles. International Journal of Computer Science and Information Technology (IJCSIT), 2(5), pp.2350-2353

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. Procedia Computer Science, 1(2), 2811-2819.

Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2011). Factorization Models for Forecasting Student Performance. EDM (pp 11-20).

Van den Bosch, A. (2011). Hidden Markov Models. In Encyclopedia of Machine Learning (pp. 493-495). Springer US.

Wang, F. and Shao, H. (2004) Effective Personalized Recommendation Based on Time-Framed Navigation Clustering and Association Mining, Science Direct, Expert Systems with Applications Vol. 27, PP. 365–377

Want, R. (2006). An introduction to RFID technology. IEEE Pervasive Computing, 5(1), 25-33.

Wedemeyer, C. A. (1973). Characteristics of Open Learning Systems.

Welsh, E. T., Wanberg, C. R., Brown, K. G., & Simmering, M. J. (2003). E- learning: emerging uses, empirical results and future directions. . International Journal of Training and Development, 7(4), 245-258, 2-3.

Wickerhauser, M. V. (1994). Large-rank approximate principal component analysis with wavelets for signal feature discrimination and the inversion of complicated maps. Journal of Chemical Information and Computer Sciences, 34(5), 1036-1046.

Wickström, G., & Bendix, T. (2000). The" Hawthorne effect"—what did the original Hawthorne studies actually show? Scandinavian journal of work, environment & health, 363-367.

Wiersma, W., and Jurs, S. G. (2005) Research methods in education: An introduction, 9th edition, USA: Pearson.

Witten, I. and Frank, E. (2005). Data mining. Amsterdam: Morgan Kaufman.

Yadav, S. K., & Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. arXiv preprint arXiv:1203.3832.

Yang, Z. (2010). Machine learning approaches to bioinformatics. Singapore: World Scientific.

Appendix A

System Development

1) The recommendation System (SPS system) contains:

1.1 Data Collection 1 (student)

The data collection process is important for the recommendation system as there are many students (about 2650) from whom the data has been collected. The performance of the recommendation system will be improved by increasing the number of relevant features used. The initial step is focussed on dividing the project into several steps, this step is focusing on applying the initial algorithm and for this steps the researcher does not need more than the defined features. The important point is that this stage represents just one of three components in the system, so in other components in the next stages the researcher will use a wider range of features to improve the accuracy.

Decision tree in machine learning

Decision tree learning; most of the machine learning uses a predictive model through the application of a decision tree. The use of decision tree concept in this report is to provide an algorithm based on ID3 concept for the machine learning (Sra, Nowozin and Wright, 2012: 1-17).

Learning styles and VAK

Developed in the 1920s for the purposes of providing different learning styles this component will be important for the development of this report. The VAK Learning Styles Model uses the following learning styles: kinaesthetic, auditory and visual. This concept will be important in the integration of the old machine learning model, the improvement of the old machine learning model and the integration of a new learning style for the purposes of describing the learning styles of the learners.

Weka

It is a Java coded containing wide range of algorithm for defining machine learning. Weka has wide variety of programming tools capable to perform visualization, association rules, clustering, regression, classification and data pre-processing; the main aim for using Weka is its ability to fit in any machine learning scheme.

The decision tree is good for understanding the problem that is important for the progress and growth of the students. There are many alternatives that will be present for the improvement of the system. The alternatives for solution of the problem have been developed for proper implementation of the solution process (Witten and Frank, 2005: 161). There are many features that are considered for the improvement of the recommendation system. The new and improved features are important for enhancing the machine learning and improvement of the analysis of the data (Rokach and Maimon, 2008: 231).

2.2 Building the ML model

After using Weka and cleaning the data the researcher generated the ID3.java code which represent a decision tree, the researcher presented the tree as a comment at the beginning of the code, after that the researcher made changes in the function classify to return the value of the expected performance. This is a good method to develop the ML model that will be recording, organizing and sharing some detailed information that will be used in the project.

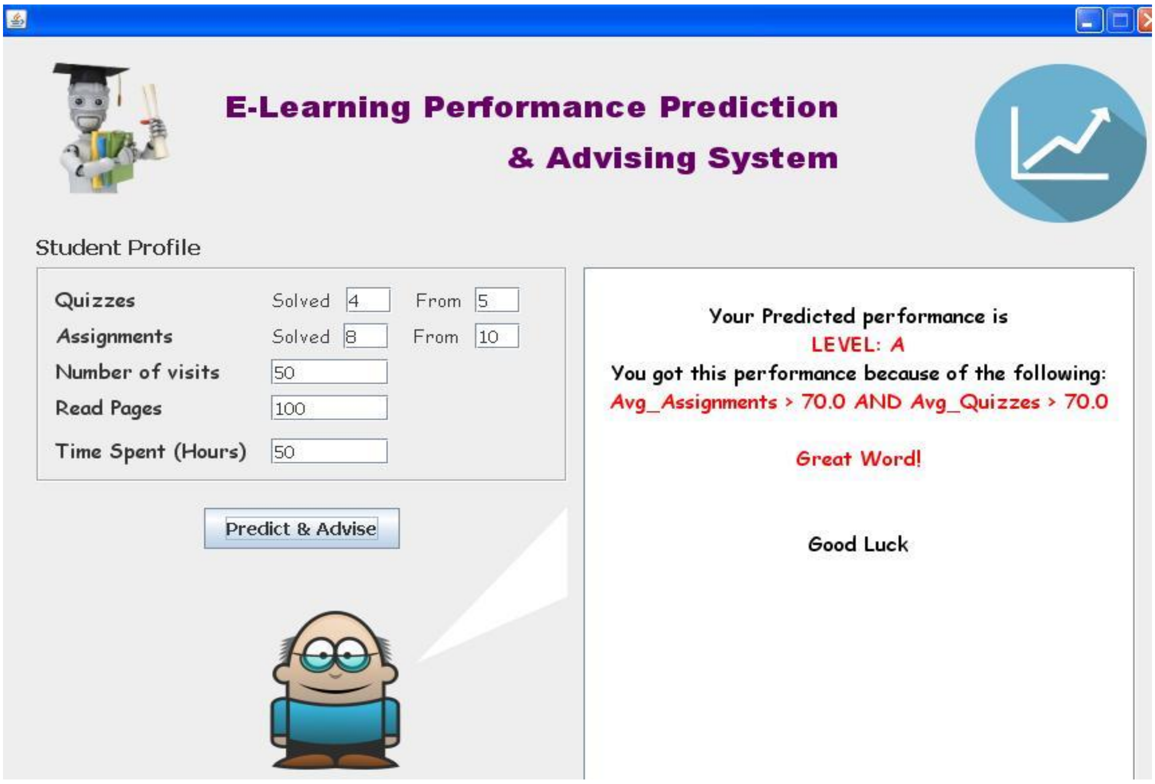
The data is split in this manner that will improve the performance. It is very much sensible for splitting the data in this way for improving the performance of the recommendation system. The decision tree and machine learning concepts are appropriately designed for ensuring that the solutions are found for different problems using the different aspects of the software. It is appropriate for the progress and success of the businesses to achieve the right steps and methods that will be good for the business. If there are certain problems that are supposed to be resolved then it will be done in the right manner without focusing on the issues or problems. If the machine learning is ensured that will be developing a strong model for the growth of the business (European Conference on Machine Learning, 2010).

Machine learning

The evolution of traditional computing to the current artificial intelligence era saw the development of computational learning theory and pattern recognition encompassing machine learning. According to Arthur Samuel the concept of machine learning describe the ability of a computer to understand learning concepts with less instruction (Yang, 2010). Generally, the concept implied in machine learning entails creating a learning algorithm with the ability to predict the outcomes of data manipulation (Baldi and Brunak, 1998: 14).

Machine learning has the capability of developing the following applications: supervised learning which entails understanding the design of the given input and producing the rule in regard to the design of the output given; unsupervised machine learning entails leaving the algorithm unsupervised to discover the set of hidden instructions and produce the desired output; and finally, reinforcement learning entails the machine learning which enables dynamic interactions which the environment for the purposes of achieving a particular objective. Such level of machine learning is demonstrated in driving.

1.3 Building the application:



E-Learning Performance Prediction
& Advising System

Student Profile

Quizzes	Solved	4	From	5
Assignments	Solved	8	From	10
Number of visits	50			
Read Pages	100			
Time Spent (Hours)	50			

Your Predicted performance is
LEVEL: A
You got this performance because of the following:
Avg_Assignments > 70.0 AND Avg_Quizzes > 70.0

Great Word!

Good Luck

Figure 20 E-learning performance prediction and advising system

The application that has been developed for collecting and analyzing the information of the students has been developed with suitable research and analysis. The basic design of the interface is actually also tied up with the structure of the data. An instructor will be able to access important information about the student performance through the application. There will be more focus towards developing standard criteria that will be assessing the performance. The features that will be needed for the full system will focus on finding aggregate or average statistical information that will guide the instructor to assess the student performance. If the basic information of the student profile is known, e.g. the read pages or number of visits, then the application will calculate the average for all the students that are enrolled showing the overall performance and improvement in the students. A simple interface is desirable as it will be focusing on the UI specific components and the functionality is much easier and having features. The right information will be used to reach effective results.

The design of the interface is very much tied in to the structure of the data. This means that if an instructor had a different dataset, they would have to recode the system to ask for input for the features that are present in this dataset. So it would be much better for the interface to be decoupled from the data and allow a more flexible approach. For example, there could be an XML document outside the interface that defines what features are required as input - in fact, these could be derived from the data itself, i.e. the program could be given a dataset as input, then it can work out what the features are and provide a way of inputting these in the GUI, and then run the classifier learner on the data automatically. In the next stages similar work will be done. The logic of the system and the UI level will be improved in such a manner that it will be a more adaptive system in the end.

2) Improving learning style (First phase):

2.1 Data Collection 2 (learning style):

Another type of data was collected. The second type of data was produced using the VAK (Visual, Auditory, and Kinesthetic Learning Styles) test.

There are six columns: the student ID, time of this test (for each student the researcher performed a number of tests periodically, so when this value equals 1 that means the 1st test, 5 means 5th test...etc.), and 3 columns for V, A, and K values which are the output of the VAK test. The last column reflects the grades of the student: has it improved, worsened, or achieved a stable case. Each row consists of the result of the VAK test (note that V+A+K=1) representing the values of Visual, Aural, and Kinesthetic depending on the VAK approach to evaluating the learning style of a student. In the last column I, S, D represents the change in the status of the student i.e. "Improved Performance" (I), "Static Performance" (S) and "Deteriorated Performance" (D).

2.2 Analysing the problem and design:

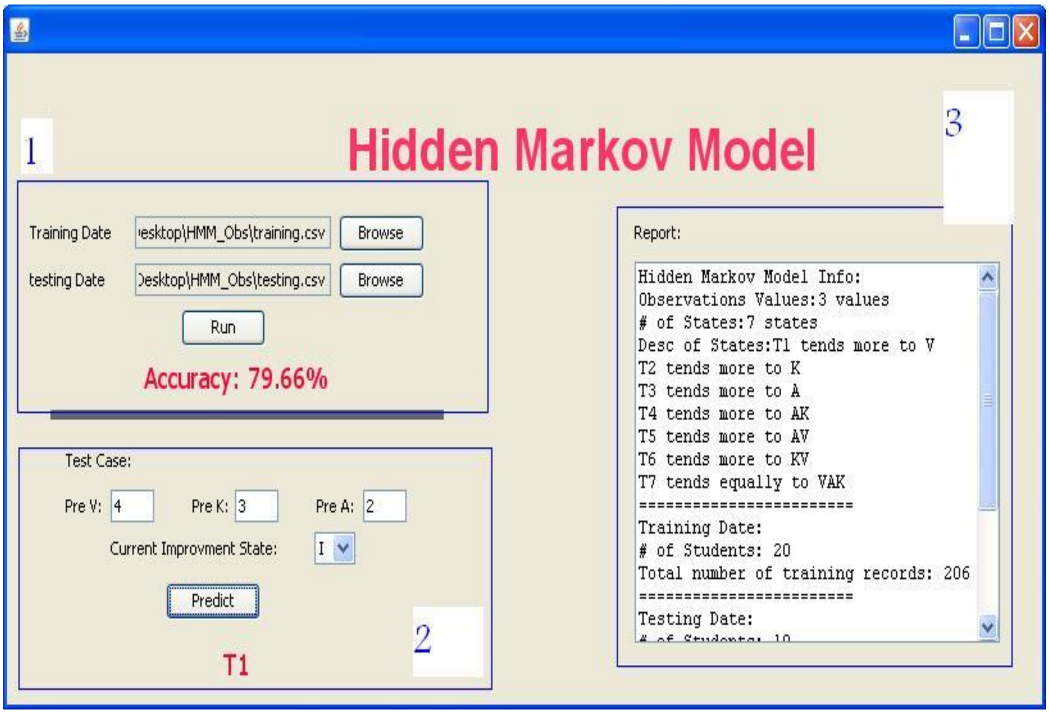
Most research shows that the students who were taught by using an adaptive learning style system performed significantly better in academic achievement than those who were taught by the same material without adaptation to the learning style.

Training data will consist of an initial test student group, through which the researcher will choose a set of students and evaluate their learning style using available online tests (such as http://www.businessballs.com/vaklearningstylestest.htm).

The researcher will select a set of students randomly from the students which the researcher have already used their data in the first stage. The researcher will check the change of this learning style after each period and record the changes. This means that each period could be one semester or one month.

2.3 Implementing the Design:

It is an implementation for the design which the researcher has worked in steps. Actually the real usage for this component would be shown clearly in next stages when all components are integrated. In other words, there is no UI for this component since it would be used internally in the next stages. For testing purposes, a simple UI was implemented to be able to test the accuracy of this specific component to be able to predict the benefit of adding it to the whole project. The system internally will train a HMM using training data and then test the model using testing data and calculate the accuracy.



Training Date: 'esktop\HMM_Obs\training.csv

testing Date: Desktop\HMM_Obs\testing.csv

Accuracy: 79.66%

Test Case:

Pre V: 4 Pre K: 3 Pre A: 2

Current Improvment State: I

T1

Report:

```
Hidden Markov Model Info:
Observations Values:3 values
# of States:7 states
Desc of States:T1 tends more to V
T2 tends more to K
T3 tends more to A
T4 tends more to AK
T5 tends more to AV
T6 tends more to KV
T7 tends equally to VAK
========================
Training Date:
# of Students: 20
Total number of training records: 206
========================
Testing Date:
# of Students: 10
```

Figure 6 Hidden Markov Model

Part 1: this provides the system with two CSV files (from the collected data in a previous step), one for training and another for testing. Then the system trains the HMM which has been described in the design and then tests this using the testing data, showing the results.

Part 2: this is a simple UI to be able to do some fast tests in the system. Again, V, A, K refers to Visual Auditory Kinesthetic learning styles (Bhat, 2004: 371-374). So the case in the figure means: if V=4 (of ten i.e. 4/10), K=3, A=2 and the student is improved then the learning style of the student is expected to be Visual. For these stages the researcher will use this information in giving more suitable suggestions to learners regarding their learning practices.

Part 3: is a report for the training operation which describes the HMM items and its results. Regarding to the accuracy: it is a good accuracy for an HMM if it is about 80% of the cases the system predicting the learning style correctly.

3) The second improvement related to the student's activities sequence contains:

3.1 Data collection 3 (activities sequence):

The main target in this stage is collecting data representing the “ontology" for E-learning concepts; the focus here is on computer science related concepts. The data has been collected from Wikipedia since it provides the concepts, their relationship and also there is a brief description for each concept which would help in matching this hierarchy with student courses names.

It is important to clarify that Wikipedia is not a source of information in this research but it is used for technical purpose (the hierarchy in wiki represents correct semantic relationship). Wikipedia was found to follow similar structure to many hardcopy encyclopedias e.g. Encyclopedia of Computer Science and Concise Encyclopedia of Computer Science. Given that Wikipedia is available online it was easier to use it to extract data to train the system to understand the relations between different concepts. The main target in this stage was collecting data representing an ontology for E-learning concepts; the researcher focused here on computer science-related concepts. The data has been gathered from Wikipedia since it was thought to provide the concepts and their relationships. Also, there is a brief explanation of each concept which would help in matching this hierarchy with student course names. In more detail, the data has been collected from Wikipedia since it provides the concepts and their relationship (via links, for example under algorithm category in Wikipedia we can see algorithms design.etc). Also there is a brief description for each concept which would help us in matching this hierarchy with student course names. Noticeably, there is no available data which provides exactly what this stage seeks to achieve; for this reason the researcher parsed Wikipedia and detect the relation between concepts using the relationship between pages. "During recent years, Wikipedia, the world's largest collaborative encyclopedia, has accumulated vast amount of semi structured knowledge (17 million concepts in total and 3 million in English), which to some extent reflects human's cognition on relationship." (Bu, et al., 2011).

The XML file includes different concepts in computer science. For example:

```
<page>

<title>Category: Algorithms</title>

<ns>14</ns>

<id>691136</id>

<revision>

<id>652967035</id>

<parentid>543792699</parentid>

<model>wikitext</model>

<format>text/x-wiki</format>
```

```
<text xml:space="preserve" bytes="253">{{catdiffuse}}

{{Commons cat|Algorithms}}

{{Cat main|Algorithm}}


[[Category:Algorithms and data structures]]

[[Category:Computer programming]]

[[Category:Mathematical concepts]]

[[Category:Mathematical problem solving]]

[[Category:Applied mathematics]]</text>

</revision>

</page>
```

This concept is related with: Algorithms and data structures, Computer programming, Mathematical concepts, Mathematical problem solving, Applied mathematics. Also there is an ID and parent ID for each concept. In this way the researcher connects between all concepts in this file.

Because of no available data which give exactly what was required, the researcher paraphrased Wikipedia and detected the relation between concepts using the relationship between pages.

The accuracy rate with this file (Wikipedia.xml); will be good because the hierarchy in Wikipedia represents correct semantic relationships. Considering the issue of other ontologies available for this online, the researcher identified available ontologies such as wordNet. The researcher checked them in detail before deciding to build an ontology because it could not give the information needed; the researcher thought the problem is found because these ontologies are generic and not focusing on computer science like the present case.

3.2 Analysing the problem and design:

While ontology is typically understood as a class hierarchy in a taxonomic sense, it can also be understood as a class definition or as simple subsumption relations. Nonetheless, ontologies do not have to be limited to such a strict form. Ontologies can also be understood outside conservative definitions that focus more on terminology and the traditional logic sense than improving knowledge regarding a subject. One of the basic problems when modelling students by ontology is what to consider and what to ignore. If the researcher has used ID3 already, which is a strong decision making system based on decision trees, and add the power of two adaptive components such as HMM and ontologies, it is possible to truly get closer to the adaptive system students and tutors need.

3.3 Implementing the design:

After the researcher have collected the required data to build the ontology of concepts related to the courses (the researcher chose CS field as a case study), in this stage two main components will be:

1- Course Concepts: This component is responsible for reading the ontology of courses and concepts and representing it using "Concept" data structure which consist of:

*The name or title of the concept, for example "Algorithms and data structures"

```
String title = "";
```

*A list of the possible categories for this concept, ex.. a possible category for "Algorithms and data structures" is "Computer Science"

```
ArrayList<String> categories = new ArrayList<>();
```

*A list of words which could describe this concept, ex.. a possible tags for "Algorithms and data structures" are "Algorithms", "Analysis of algorithm", "Abstract data types"..etc

```
ArrayList<String> tags = new ArrayList<>();
```

The last two lists are temporary lists which would be used to build the following two lists

*parent concepts or "Hyponymous relationships": between the concept and its "parents", this relation is also called is-a relation since if 'A' is parent concept for 'B', so 'B' is-a 'A', ex: "Analysis of algorithm" is-a "Algorithms and data structures", and "Algorithms and data structures" is-a "Computer Science"

```
ArrayList<Concept> ParentConcept = new ArrayList<>();
```

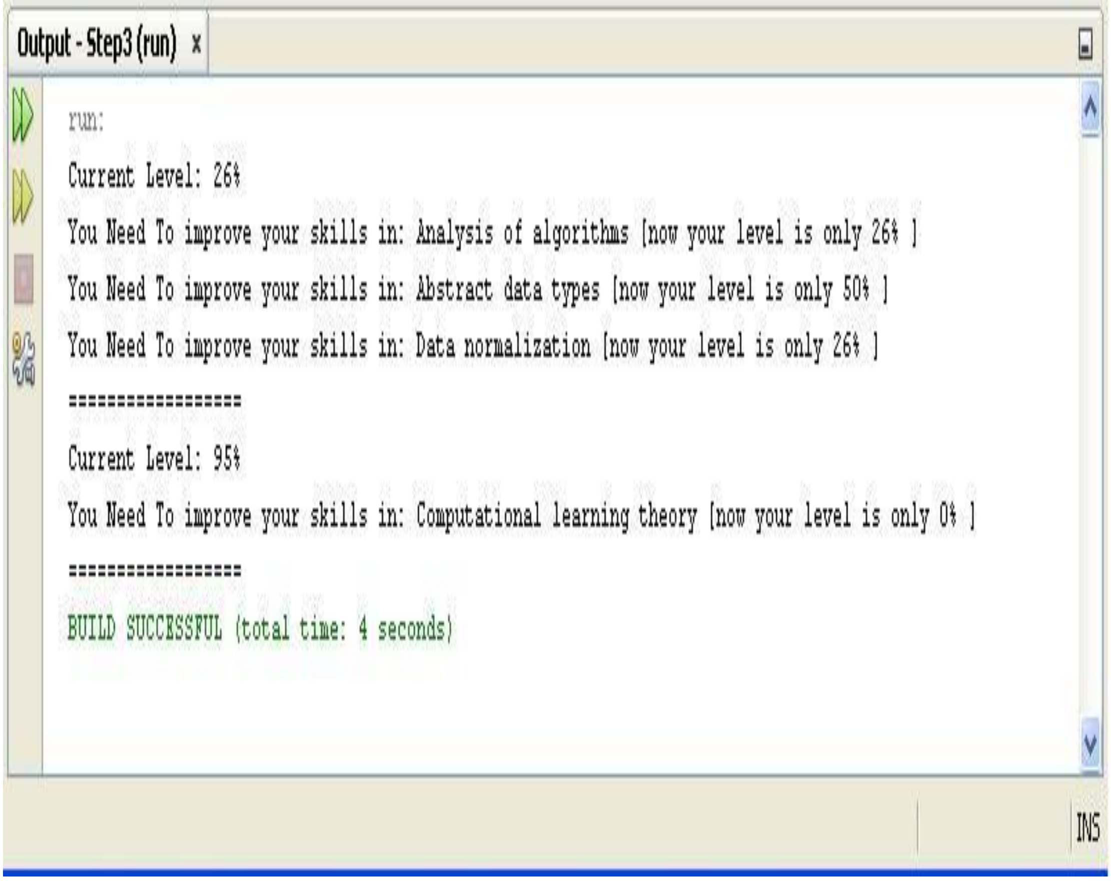
*child concept or has-part relationship: between the concept and its sub-concepts, it is the opposite of the last case.. ex: "Computer science" has-part "Algorithms and data structures"

ArrayList<Concept> ChildConcept = new ArrayList<>();

*Finally, a value representing an evaluation for the knowledge of a person (a student) in this concept, it take its values between [0-1], 1 means he completely understand this concept, 0 means the opposite.

double known = 0;

2- Student Concepts Model Builder which will represent the knowledge of the students using the defined ontology. This component also is responsible for extending the known information about the students by applying reasoning rules on the available information about the student.



```
run:
Current Level: 26%
You Need To improve your skills in: Analysis of algorithms [now your level is only 26% ]
You Need To improve your skills in: Abstract data types [now your level is only 50% ]
You Need To improve your skills in: Data normalization [now your level is only 26% ]
==================
Current Level: 95%
You Need To improve your skills in: Computational learning theory [now your level is only 0% ]
==================
BUILD SUCCESSFUL (total time: 4 seconds)
```

Figure 7 Program output

In the program the output is:

Current Level: 26%

```
You need to improve your skills in: Analysis of algorithms [now your level is only 26%]

You need to improve your skills in: Abstract data types [now your level is only 50%]

You need to improve your skills in: Data normalization [now your level is only 26%]

==================

Current Level: 95%

You need to improve your skills in: Computational learning theory [now your level is only 0%]

==================
```

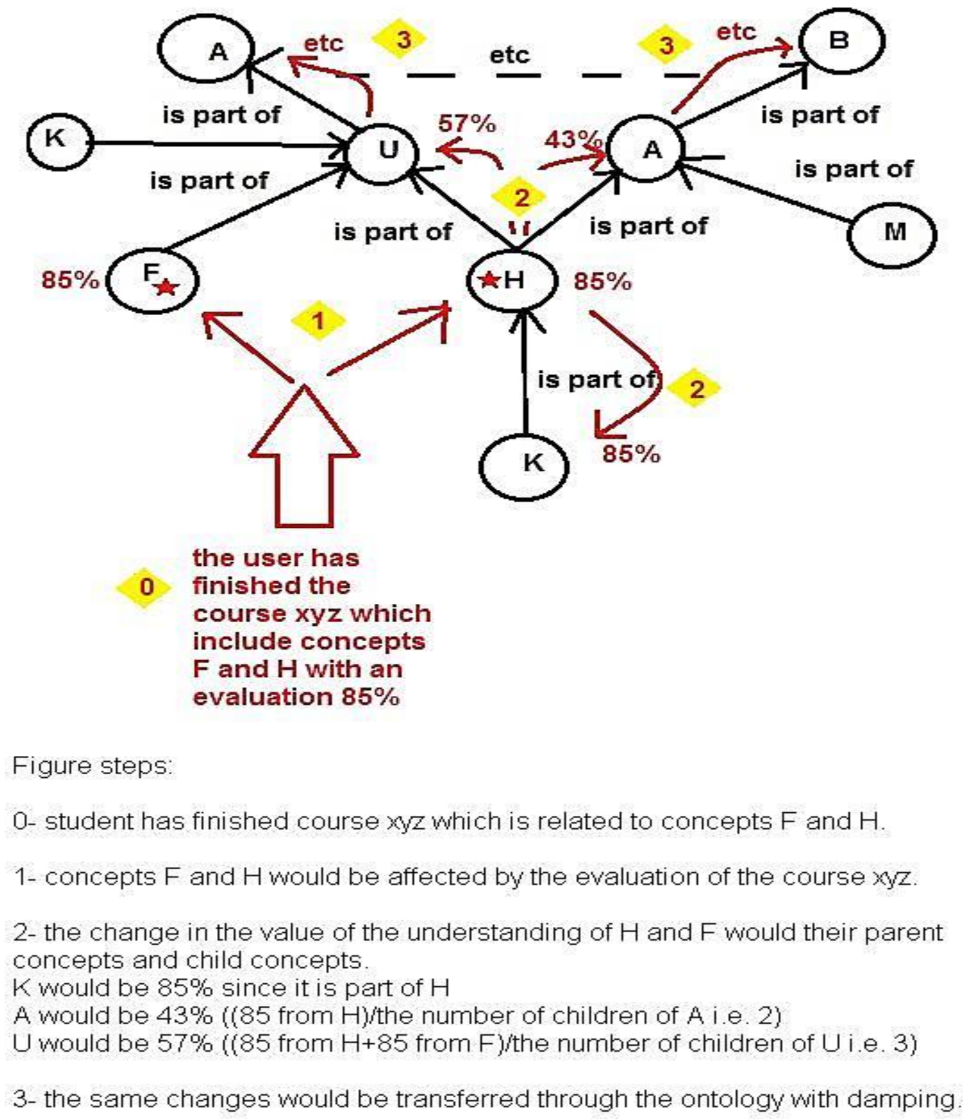
There are two percentages in the current level (26%) and (95%) these two numbers are for two different requests. For this example, two tests cases were added:

```
 String report1 = john. Diagnose("Algorithms and data structures");
System.out.println(report1);
 String report2 = john.Diagnose("Machine learning");
System.out.println(report2);
```

The researcher asked the system to check the level of ‘john’ in "Algorithms and data structures" and in "Machine learning". So the first number is for "Algorithms and data structures" and the second number is for "Machine learning". It would help in diagnosing the reason of the bad results and being able to give suggestions to improve it.

In this current step, when students have a problem in the concept X, the system would conclude (using the ontology) that concept X could not be understood in a perfect way without improving the other concepts: A, B, C for example, Then the system evaluates the level of the student in A, B, C to conclude the reasons of the bad result in X. Let us say it evaluated the understanding of the student in A: 85%, B: 12%, C: 32%. Then the system will find that the problem is in B, C and will suggest improving them.

Concerning the diagnosis and generation of results (e.g. "Current Level: 26% you need to improve your skills in: Analysis of algorithms), the system derives these values from its previous experience with this use. After finishing each course, the evaluation of the student would be entered to the system, for example in course H the system evaluation was 75%. Then the system fills related concepts (not only H but also related concepts) using this value (for example H consist of concepts A and U) then the values of H, A, and U would be increased. And the same for related concepts to A and U (since its value would be changed, it related concepts would be affected).



the user has finished the course xyz which include concepts F and H with an evaluation 85%

Figure steps:

0- student has finished course xyz which is related to concepts F and H.

1- concepts F and H would be affected by the evaluation of the course xyz.

2- the change in the value of the understanding of H and F would their parent concepts and child concepts.
K would be 85% since it is part of H
A would be 43% ((85 from H)/the number of children of A i.e. 2)
U would be 57% ((85 from H+85 from F)/the number of children of U i.e. 3)

3- the same changes would be transferred through the ontology with damping.

Figure 8 Score calculation process

The student's score for A would improve by 43% from its original value, in our example the initial score was 0 but in general it would improve 43%. The student achieved 85%, so the question is why the system thinks that the student needs to improve his performance by 43%? Until this step the system has not given any suggestion it is just the training stage.

4) Integrating the two systems:

The objective is to develop an improved machine learning model which incorporates the traditional ID3 technique and the Hidden Markov model (which is the SPS system) plus the first improvement of the SPS system machine learning model to include learning styles and the second improvement to incorporate the student's activities sequence. Consequently, the following models were developed.

The integration of the two systems should result in good performance of the overall system. The work will be done in a more efficient and effective manner as the right attributes from both systems will be combined to make the final work stronger and effective. The long run successful completion of the learning activities will be ensured by focusing on the right strategies and focusing on the growth of the students.

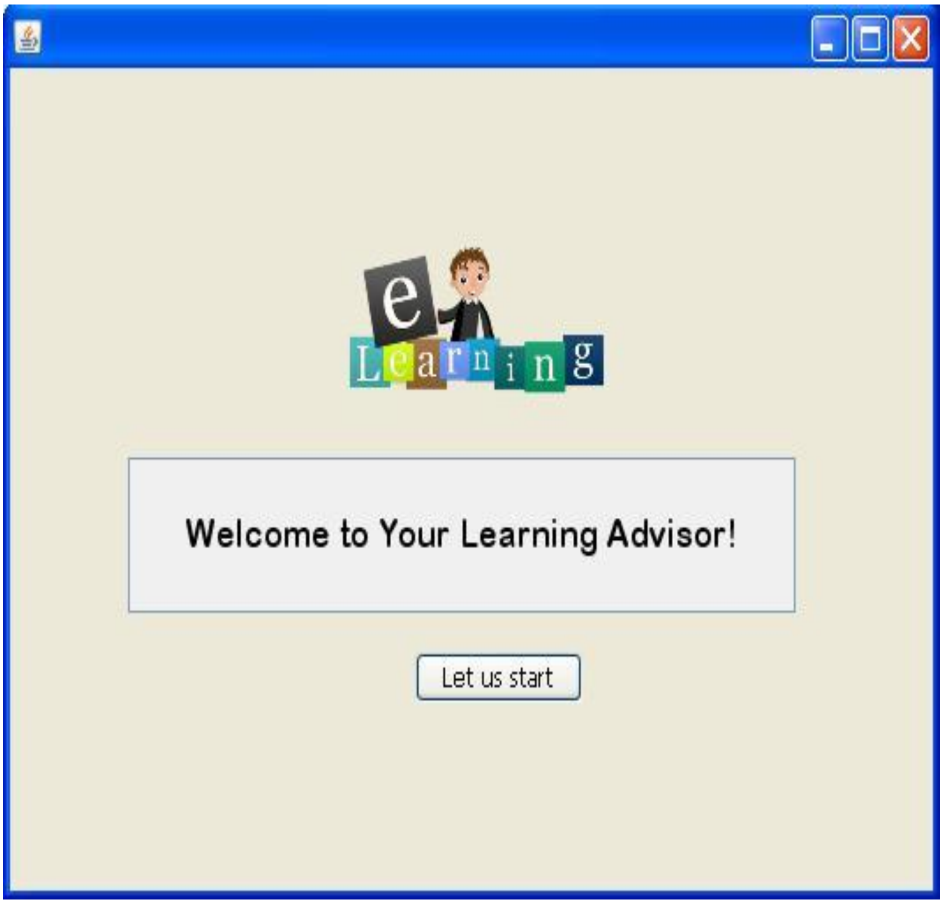
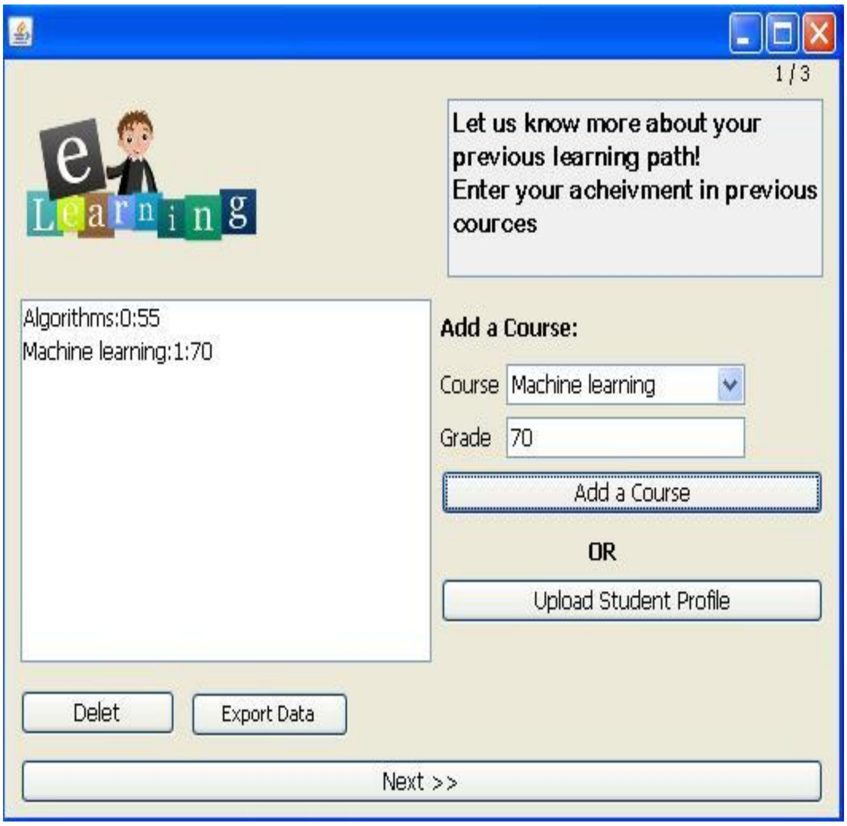

Welcome to Your Learning Advisor!

Figure 9 Welcome screen

The first phase of the developed machine leaning model welcomes the learner to the model. The model has the opening clause 'Welcome to Your Learning Advisor". This assures the students that she or he is in the right place for his learning problems. This interface initializes the learning sequence of the learner.



1 / 3

Let us know more about your previous learning path!
Enter your acheivment in previous cources

Algorithms:0:55
Machine learning:1:70

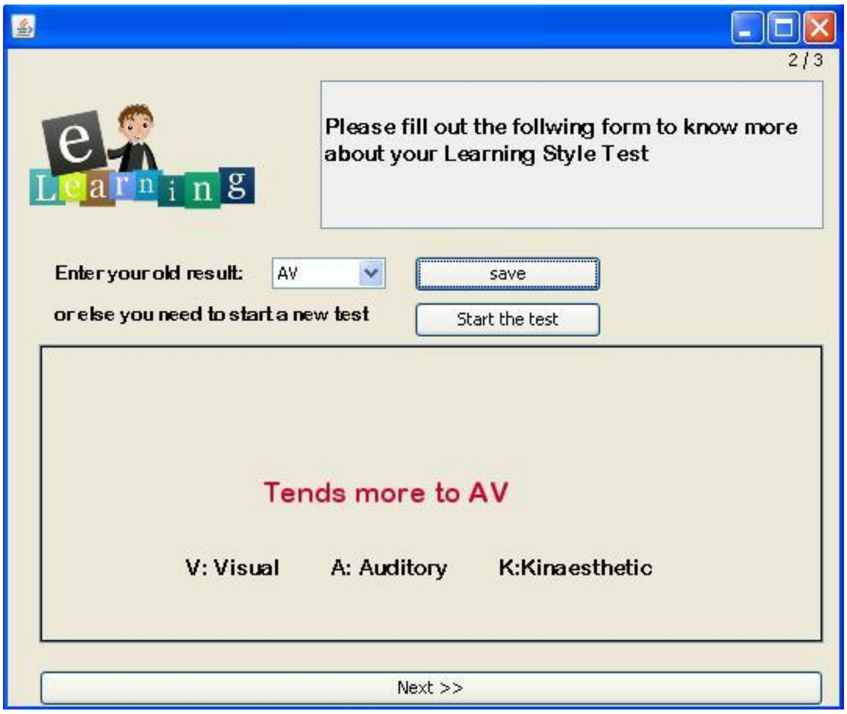
Add a Course:

Course: Machine learning

Grade: 70

Figure 10 Information about the previous learning path screen

After the learner logs in to the improved learning model, the system requires the learner to provide information about the previous learning path. This is crucial for the machine to integrate the student information and understand better how to incorporate the student profiles.



Please fill out the follwing form to know more about your Learning Style Test

Enter your old result: AV

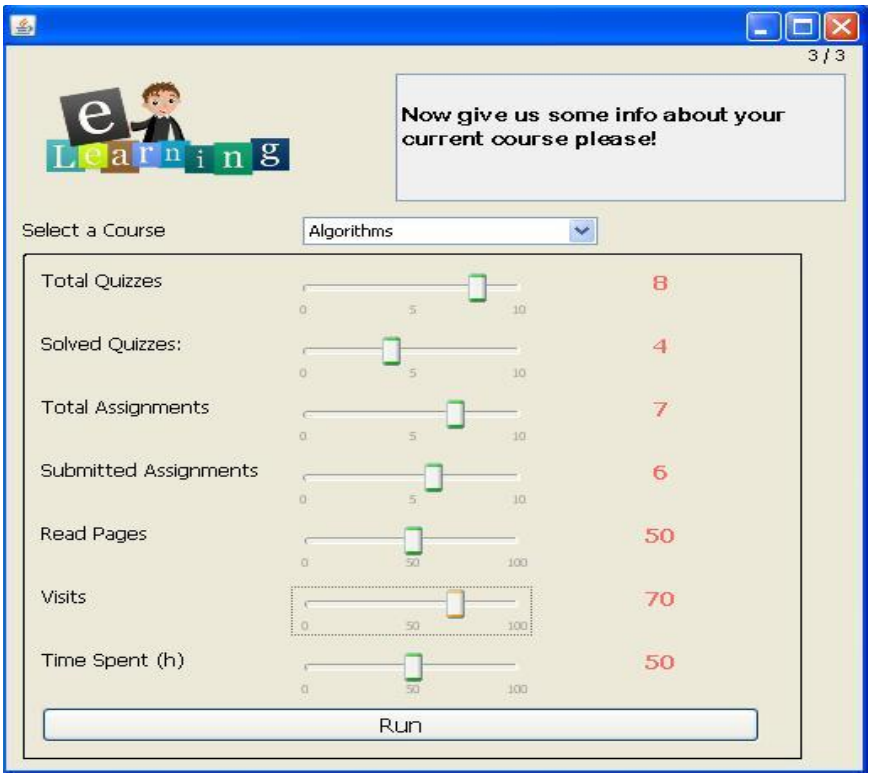
or else you need to start a new test

Tends more to AV

V: Visual A: Auditory K:Kinaesthetic

Figure 11 Learning style prediction screen

The model provides the learner about his or her learning style test. Here the student is supposed to undergo a test in order to understand his learning style. The model feeds in the test scores of the learner and then predicts the learning style of the learner; this can be Visual, Auditory and Kinaesthetic. This is crucial to understand the different learning styles for the learner as that determines the progress and control of the activities. The test is developed by focusing on the end user skills and assessing their overall performance.



3 / 3

Now give us some info about your current course please!

Select a Course: Algorithms

Total Quizzes 8

Solved Quizzes: 4

Total Assignments 7

Submitted Assignments 6

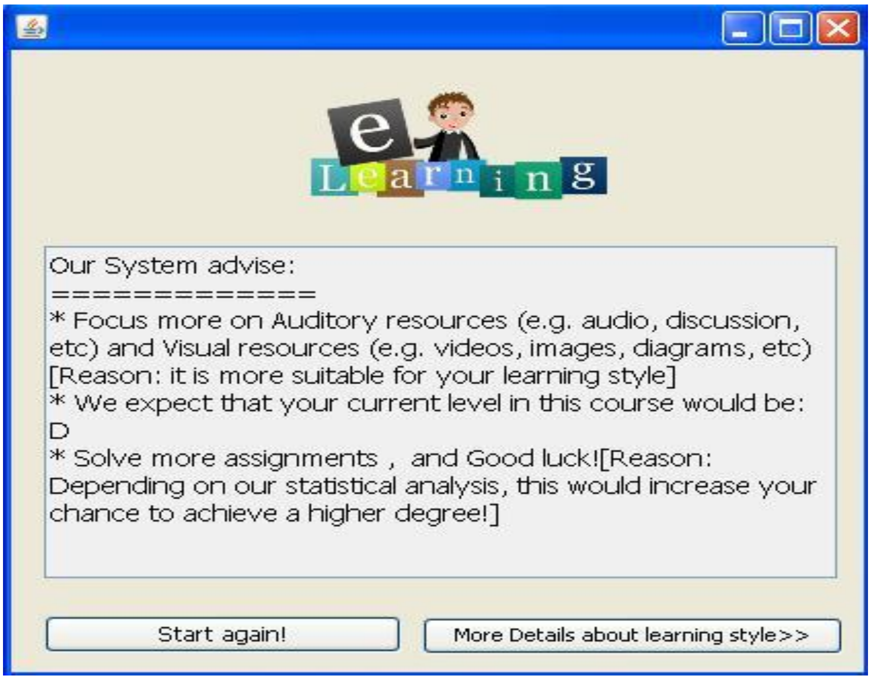
Read Pages 50

Visits 70

Time Spent (h) 50

Figure 12 Student's current course screen

The model intends to seek more information concerning the student's current course. This is applied through outlining all the activities of the students undergoing a particular course. For example, the total quizzes that the student was provided, the total answered, the total number of assignments and the total number of the assignments that the student was able to submit, etc. Here the machine feeds the information to predict the student activities sequence. This is very much an informative place where the user can view the previous record and an overall summary of the performance. The information is shown to clarify the progress or improvements in the performance.



Our System advise:
==============
* Focus more on Auditory resources (e.g. audio, discussion, etc) and Visual resources (e.g. videos, images, diagrams, etc) [Reason: it is more suitable for your learning style]
* We expect that your current level in this course would be:
D
* Solve more assignments , and Good luck![Reason: Depending on our statistical analysis, this would increase your chance to achieve a higher degree!]

Figure 13 System advise screen

After the machine learning model incorporates all the student learning style and leaning activities sequence, it then provides advisory information to the student

The system has been trained using data which was collected previously by the researcher. This data can be changed if required, which will cause a recalculation of the transition probabilities.

The data from for the transition probabilities was obtained from asking a set of students to do VAK test a number of times periodically in the one semester and relating the results with their grades.

Appendix B

Popular approaches

Introduction

In this section, the researcher will outline the popular approaches used by other researchers in the area. It is important to review the review approaches to understand how similar problems were solved in the past. This will then better inform the development of the system.

The approaches

Pardos et al. (2007) studied the use of granularity (the number of modelled skills) for the prediction of student performance. Skills can be of any category; if one considered geometry, skills could be: congruence, equation solving, area of a triangle, volume of a sphere and many more. The number of modelled skills in the study was set to different values: 1, 5, 39 and 106. In the end, when they compared final results, the optimal granularity among the four presented values was 39 in the case of MCAS testing – which is related to mathematics – but 106 was optimal for internal use. Internal use meant predicting the accuracy of student answers to the questions found on the online system of the university. This approach did not only rely on granularity, which deals with the modelling itself, but also on Bayesian Networks. The construction of the Bayesian Network included three layers, being: the knowledge layer, the intermediary layer with logical AND gates, and a question node layer. Thus, for the prediction, the probability was retrieved from the built network. The term 'skill model' in the study means a set of skills with the linked questions. Pardos et al. did not specifically mention all the studied skill sets. It is obvious that they analysed skills of students in different subject matters, which has relevance to this thesis. The granularity of the skills was also discussed, which stands at the base of the modelling.

Bayesian Networks are successfully used by Ramaswami et al. (2012). It is interesting that Ramaswami used granularity, but not based on the presented work of Pardos et al. (2007), but based on a presentation of the same authors, in 2006 (just one year before), with the title 'Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks'. The fine-grained breakdown approach of Ramaswami aims at classifying student marks in different categories. The research conducted used four different classifications: 2-class, 3-class, 5-class and 7-class. These values provide more accurate breakdowns, with more categories for students' marks. The scholars used an estimator algorithm with various search algorithms, such as Repeated Hill Climbing, LAGD Hill Climbing, K2, TabuSearch, Network Augmented with Tree, and Hill Climbing and observed the results. In conclusion, from the above algorithms it was found that the most accurate was Network Augmented with Tree. The scholars suggested the accuracy for the Network Augmented with Tree algorithm based on comparing multiple results. As stated, they have used six different search algorithms on four different mark classifications. Comparing the table of results, they have reached the conclusion of the highest overall predicting accuracy obtained by the tested BN model, using the chosen algorithms.

Table 11 Bayesian Network Models prediction accuracy (%) presented by Ramaswami et al. (2012)

Search Algorithm	**Predictive Accuracy**			
	2-class (IG-9)	**3-class (IG-13)**	**5-class (IG-19)**	**7-class (IG-23)**
HillClimber (HC)	82.2575	59.1647	40.214	31.4981
K2	82.2575	59.1647	40.214	31.0321
LAGDHillClimber (LC)	84.5713	59.7169	40.7318	33.8799
RepeatedHillClimber (RC)	82.2575	59.1647	40.7318	31.4981
TabuSearch (TS)	84.8809	60.5627	52.1401	31.4636
TAN (TN)	84.9154	63.3069	52.1401	42.3196

Table 11 shows the prediction accuracy, as presented with the fine-granularity on marks. A graphical representation of the same can be found in Figure 29.

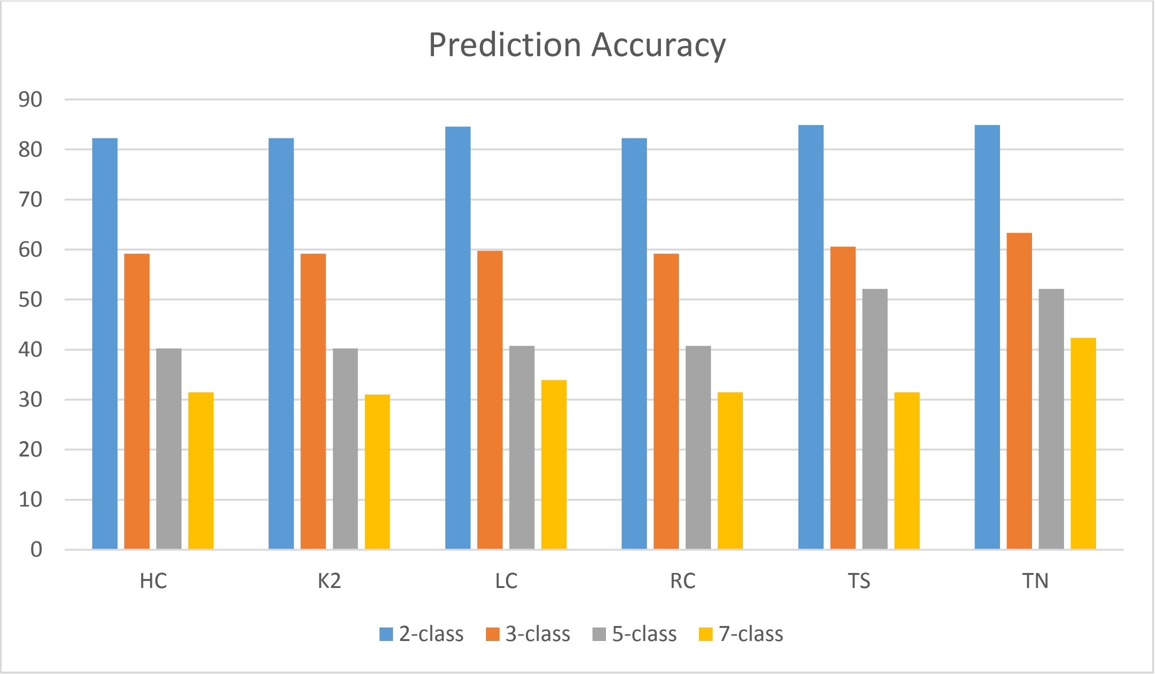


Figure 29 Prediction Accuracy of Bayesian Network Model presented by Ramaswami et al. (2012)

Adhatrao et al. (2013) used ID3 and C4.5 classification algorithms. C4.5 is a newer algorithm derived from ID3 and is stronger and more accurate in certain cases. Using both of the algorithms in the model resulted in increased accuracy. Thus, based on information collected from students about their background and past study results, the researcher can predict which students will probably become excellent students. Similarly, students with chance of failure must receive more support from the teachers.

Baradwaj & Pal (2012) also used ID3 to predict student performance after they have attended their examinations. They focused on data mining techniques specifically used to extract relevant data, which allows better analysis. The researcher can observe potential failure and students who may want to leave higher education. Knowing the possible outcome, teachers can handle situations and avoid future dropouts. Data was collected from the VBS Purvanchal University, located in Jaunpur (India). After the data had been collected, there was a selection process intended to keep only relevant information. For each student, multiple variables were stored, such as previous semester marks, seminar performance, general proficiency, end semester marks, attendance, lab work, class test grade, and assignment.

In conclusion, the research is helpful to discern between students who will not fail and those who will probably fail. Thus, both students and teachers will be able to handle needs accordingly, which means teachers can help the students that might fail based on the prediction.

While student performance is analysed, there is a possibility of other factors that would influence the final performance in addition to the ones selected and used in this research. Personal motivation, social background and various other factors can determine – to a certain degree which is yet to be analysed in detail – how much students will work for performing better or not. Thus, helping the students would be improved if more information would be included in the analysis.

Yadav & Pal (2012) wanted to target Engineering students and improve their academic results. To obtain the improvement, a very accurate prediction is required. Proving the results was only possible after the next examination session was finished and marks were available. They used the ID3, C4.5 and CART algorithms. Students who were predicted by these decision tree algorithms to fail had a chance to improve results and to study more. The results also proved the method efficient, and proved that Engineering students can use such predictive software to measure their own quality of studies.

As stated by Beal et al. (2007), Hidden Markov Models are also powerful when predicting students' learning patterns. They have used HMMs to measure the engagement of each student, classifying three different levels of engagement. In the conclusion of the study, it is shown that intelligent learning systems can use implementations of HMMs to predict results better, thus helping the students more. Engagement was measured and classified, and the algorithm was able to prove useful for mathematics, the subject on which this study was conducted.

Thai-Nghe et al. (2011) proposed a new approach which uses factorisation models for predicting student performance. The basis of the mathematical approach presented in this research is that as time passes, the knowledge of any student will improve and cumulate. Thus, a factorisation model is efficient to predict performance. The tensor factorisation introduced in this article can be improved and taken to higher levels of accuracy. The next levels of accuracy can be reached with more complex prediction algorithms in combination with factorisation models. Tensor factorisation can also be extended to propose exercises to the students.

Appendix C

Publications

One publication has been produced during the PhD project. Below is the resultant publication:

1. Alghamdi, F. and Jensen, R., (2014). Latest Trends in Data Mining for E-learning Systems. *International Journal of Information Technology & Computer Science*, 17 (2), 49-60.

Appendix D

Learning styles questionnaire

The researcher designed a questionnaire which contains several questions about learner personality, attitude and behaviour. Based on many cognitive research this questionnaire consists of 30 questions, each of them has three different answers and each of them reflect certain learning styles.

The researcher used different information sources available on the internet to collect questions that can be used to determine the learning style of each student. All sources used to obtain the questions were previously used in other studies. This reduced the effort and time needed to conduct this step as all question used were previously tested. A list of information sources used to obtain the questions are listed below:

1. The Everett-Hall Professional and Leadership Development Toolkit For High Impact HIV Prevention
   http://www.etr.org/cisp/assets/File/Professional-and-Leadership-Development-Toolkit.pdf
2. VARK Learning Styles Self-Assessment Questionnaire
   https://www.fitcollege.com.au/admin/InductionFiles/Files/29_V.A.R.K_Assessment_-_Which_learning_style_are_you.pdf
3. VAK Learning Styles Self-Assessment Questionnaire
   www.businessballs.com/freematerialsinword/vaklearningstylesquestionnaireselftest.doc
4. What's Your Learning Style?
   http://www.compasstoolkit.ox.ac.uk/wp-content/uploads/2015/11/Learning-Style-Quiz-Individual-Activity.pdf
5. VAK (visual-auditory-kinesthetic) learning style indicators
   http://www.career-guide.eu/uploads/vak-learning_test.pdf