A Major Project Report on

“Role of Data Mining in Educational and eLearning Systems”

Submitted for the award of the degree of

Master of Business Administration (Executive)

By

Dinesh Kumar Mandal

Roll No.: 2K13/MBA/504 | Registration No.: DTU/13/MBA/504

Under the Supervision of

Mr. Satish Dubey
Project Director (eMigrate) and
Local Ethics Counsellor
Tata Consultancy Services

Professor Dr. Pradeep Kumar Suri
Delhi School of Management,
Delhi Technological University

DELHI SCHOOL OF MANAGEMENT

DELHI TECHNOLOGICAL UNIVERSITY

Shahbad-Daulatpur, Bawana Road, Delhi -110042

May, 2015

CERTIFICATE

This is to certify that the project report entitled **“Role of Data Mining in Educational and eLearning Systems”** has been successfully completed using literature reviews, statistical techniques and advanced data mining tools by Mr. Dinesh Kumar Mandal – Roll No: 2K13/MBA/504 student of **Delhi School of Management, Delhi Technological University**.

This is further certified that this project work is a record of bona fide work done by him under my guidance. The matter embodied in this report has not been submitted in part or full to this or any other university as part of project work to the best of our knowledge.

Signature:
………………………………..
Mr. Satish Dubey
Project Director (eMigrate) and
Local Ethics Counsellor
Tata Consultancy Services

Date: …………………

Signature:
………………………………..
Professor Dr. Pradeep Kumar Suri
Delhi School of Management,
Delhi Technological University
Bawana Road, Delhi-110042

Date: …………………

DECLARATION

I, **Dinesh Kumar Mandal** Roll No: 2K13/MBA/504 student of **Delhi School of Management, Delhi Technological University**, would like to state that the project titled **"Role of Data Mining in Educational and eLearning Systems"** is an authenticated work carried out by me using existing literatures reviews, secondary data of some eLearning systems, statistical techniques, advanced data mining tools, and news and research journals available on public domains under the supervision of **Mr. Satish Dubey, Esteemed Visiting Faculty** and **Prof Dr. Pradeep Kumar Suri, HOD** in **Delhi School of Management, Delhi Technological University** for the award of the degree of **"Master of Business Administration (Executive)"** and have submitted a satisfactory report on the project. The matter embodied in this report has not been submitted in part or full to this or any other university as part of project work to the best of my knowledge.

……………………………………………

Dinesh Kumar Mandal
Roll No.: 2K13/MBA/504
Delhi School of Management,
Delhi Technological University,
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I have taken sincere efforts and interests in this project. However, it would not have been possible without the kind support and help of many individuals, organizations, and their invaluable research and ideas available in public domain. I would like to extend my sincere thanks to all of them.

I am highly indebted to **Mr. Satish Dubey, Esteemed Visiting Faculty** and **Prof Dr. Pradeep Kumar Suri, HOD** in **Delhi School of Management, Delhi Technological University** for their guidance and constant supervision as well as for suggesting areas of improvements in the project, and also for their support in completing the project.

I would like to express my gratitude towards my parents, family and faculties of **Delhi School of Management, Delhi Technological University** for their kind co-operation and encouragement which helped me in completion of this project.

I would like to express my special gratitude and thanks to my current organization **Excelsoft Technologies Private Limited** for giving me such attention and time.

My thanks and appreciations also go to my colleagues and people who have willingly helped me out with their abilities.

…………………………………………

Dinesh Kumar Mandal
Roll No.: 2K13/MBA/504
Delhi School of Management,
Delhi Technological University,
Bawana Road, Delhi-110042

List of Figures

Figures	Page No
Figure 1: Components of a typical eLearning System	7
Figure 2: Iterative nature of a data mining process	10
Figure 3: Improvement cycle in eLearning process of adopting data mining discipline	14
Figure 4: Correlation between engagement level and outcomes	22
Figure 5: Creating clusters in MS Excel 2010 and MS Analysis Services 2008 R2	27
Figure 6: Cluster profiles	27
Figure 7: Cluster diagram showing strong link between Cluster 1 and Cluster 2	28

List of Tables

Tables	Page No
Table 1: Scenario of an eLearning System	17
Table 2: Data Mining Tools and Services	19
Table 3: Business Problem 1	20
Table 4: Grading results of students scored in a test conducted for a chapter in the course	20
Table 5: Business result of correlation between engagement level and outcomes	22
Table 6: Regression output generated in MS Excel BI Tool	23
Table 7: Point prediction output generated in MS Excel BI Tool	24
Table 8: Business result of point prediction	25
Table 9: Business Problem 2	25
Table 10: Average grades chapter wise	26
Table 11: Business result of cluster analysis	29

TABLE OF CONTENTS

1 Abstract 1

2 Introduction 2

2.1 Objectives 3

2.2 Scope 3

2.3 Not in Scope 3

3 Methodology 4

4 Literature Review 5

5 Educational and eLearning Systems 6

6 The Future of eLearning 8

7 What is Data Mining 9

8 Data Mining Process 10

8.1 Problem Definition 10

8.2 Data Gathering and Preparation 11

8.3 Model Building and Evaluation 11

8.4 Knowledge Deployment 12

9 Why Data Mining 13

10 How Does This Process can Work in eLearning Systems 14

10.1 Step 1 Platform Engagement 15

10.2 Step 2 Capture Engagement Data 15

10.3 Step 3 Data Mining 15

10.4 Step 4 Presentation of Data Mining Results 15

10.5 Step 5 Improve Personalized Digital Experience 15

10.6 Step 6 Evaluate and Improvise Platform Services 15

11 Data Mining Models, Results and Interpretations 17

11.1 Business Scenario 17

11.2 Data Mining Techniques 18

1.2 Cluster Analysis 18

2.2 Regression Analysis 19

11.3 Data Mining Tools 19

11.4 Business Problems 20

1.4 Regression Analysis: Data Model 20
2.4 Cluster Analysis: Data Model 26
12 Data Mining: What it cannot do 30
13 Costs and Challenges of Data Mining Discipline 31
13.1 Lack of Business Questions 31
13.2 Costs to store logged data 31
13.3 Costs to adopt emerging standards of engagement data and communication 31
13.4 Costs to hire data systems professionals 32
13.5 Data systems interoperability 32
13.6 Creating one simplified picture from data can be difficult 33
13.7 Privacy and ethics need to be upheld 33
13.8 Choosing data mining model and appropriate algorithms 33
14 Conclusion and Future Work 34
15 Bibliography 35
16 References 36
17 Notes 37

1 ABSTRACT

eLearning is becoming a big business, with huge investments in IT Technology supporting online medium of teaching, learning, content sharing, tests and assessments, grading and evaluation of courses and programs offered by various institutes and universities. Accordingly eLearning systems have evolved a lot to offer personalized teaching and learning experience to both instructors and students through digital experience. Schools and Universities in developed countries have been pioneer in adopting eLearning systems to complement their traditional methods of teaching and evaluation. Schools and Universities in developing countries have also started eLearning technology into their educational systems to bring quality education and increase their reach among wider audience.

Service providers of eLearning systems that have significant amount of historical data about delivery and engagement of digital contents can use hidden knowledge discovered through data mining techniques to improve eLearning design, development and business strategies.

This project paper intends to highlight how data mining techniques can be integrated with the eLearning Technology to benefit instructors, students and service providers of Educational and eLearning Systems.

2 INTRODUCTION

eLearning is the use of electronic technology in learning and teaching. eLearning systems are witnessing a wider adoption as digital platforms in schools, universities and corporate houses to provide personalized means of teaching and learning to increase effectiveness of courses, programs, learning materials and teaching methodology.

eLearning systems capture vital pieces of information when learners are engaged in their platforms to access courses, chapters, learning materials, assessments, taking tests, communicating with instructors, participating in discussion boards, etc. As learners keep revisiting these platforms to complete the courses and programs, learning behaviours of each learner and effectiveness of each such digital content also get recorded over a time.

Data captured like identifiers of courses, chapters, assessments and learning materials, time duration spent on them over the period of programs, and the outcomes of each such engagement as recorded in grade books can become vital piece of information to evaluate learning behaviours, effectiveness of content packaging and services, evaluate programs, performance of learners with varying experience, skillsets, interests, educational and financial backgrounds.

Data mining techniques can help uncover hidden knowledge about all these matrices when statistical data analysis like methodology applied together with tools and techniques of information technology. Using this knowledge, schools and universities can improve their services to retain students, impart knowledge effectively, and help students increase their grades.

Companies offering eLearning services and products now understand business value of engagement knowledge in improving their own products' mix and show casing their technical capability in the educational markets.

2.1 OBJECTIVES

- To show how various data mining techniques can help derive insights and subtle patterns of knowledge hidden in piles of data of educational and eLearning systems

2.2 SCOPE

- To show how various data mining techniques can help benefit educational and eLearning systems using some key attributes of eLearning business entities

2.3 NOT IN SCOPE

- Explaining 'how-to' techniques of creating data mining models
- Explaining 'how-to' preparation of data mining structures
- Mathematical derivation of statistical methods to validate data analysis results
- Running BI Tools and SQL Data Add-ins

3 METHODOLOGY

The project study work on the subject was carried on using existing literatures reviews, secondary data of some eLearning systems, and news and research journals available on public domains.

The author himself has been working in an eLearning company for various clients like Pearson, Jones & Barltlett Learning, BYU Idaho, etc. for more than six years. These clients have been dominant players in the digital educational domain especially in the North and Central America. The experience of providing solutions and supporting their business roadmaps has been invaluable in writing this paper.

The methodology in this project study work uses the traditional statistical methods of Regression and Clustering to develop mathematical models of key attributes in eLearning business entities, and run through the possible algorithms to produce data patterns of significance.

4 LITERATURE REVIEW

A literature review is the study of a body of texts touching upon the topic on which this research study is to be carried out that aims to review the critical points of current knowledge and secondary sources (IJRCM June 2011).

Few literature reviews also have been referenced to gain insights of the topics to develop sound understanding and test models to suit this paper.

Improving Student's Performance Using Data Clustering and Neural Networks in Foreign Language Based Higher Education, The Research Bulletin of Jordan ACM, Vol. II (III): The research paper explains about how Neural Networks and Cluster like models helped learning behavior and teaching methodology in Lebanon like non English speaking country in the higher education of their native students.

Mining Students Data to Analyze Learning Behavior: A Case Study, Department of Computer Science, Islamic University of Gaza, Palestine [Alaa El-Halees]: The case study shows how techniques of machine learning helped design course and evaluation system to improve students' performance over the time.

5 EDUCATIONAL AND ELEARNING SYSTEMS

eLearning also known as Learning Management System (LMS) is a software application for the administration, documentation, tracking, reporting and delivery of electronic educational technology education courses or training programs.

LMSs range from systems for managing training and educational records to software for distributing online or blended/hybrid college courses over the Internet with features for online collaboration and personalized learning. Colleges and universities use LMSs to deliver online courses and augment on-campus courses. Corporate training departments use LMSs to deliver online training.

A typical eLearning system is shown in a diagram below. eLearning portal is the common interface for both tutors and learners to collaborate online for teaching, learning, content sharing, receiving grades and feedbacks, etc. The portal may consume specialized services, tools and contents from other third party providers as per the curriculum design.

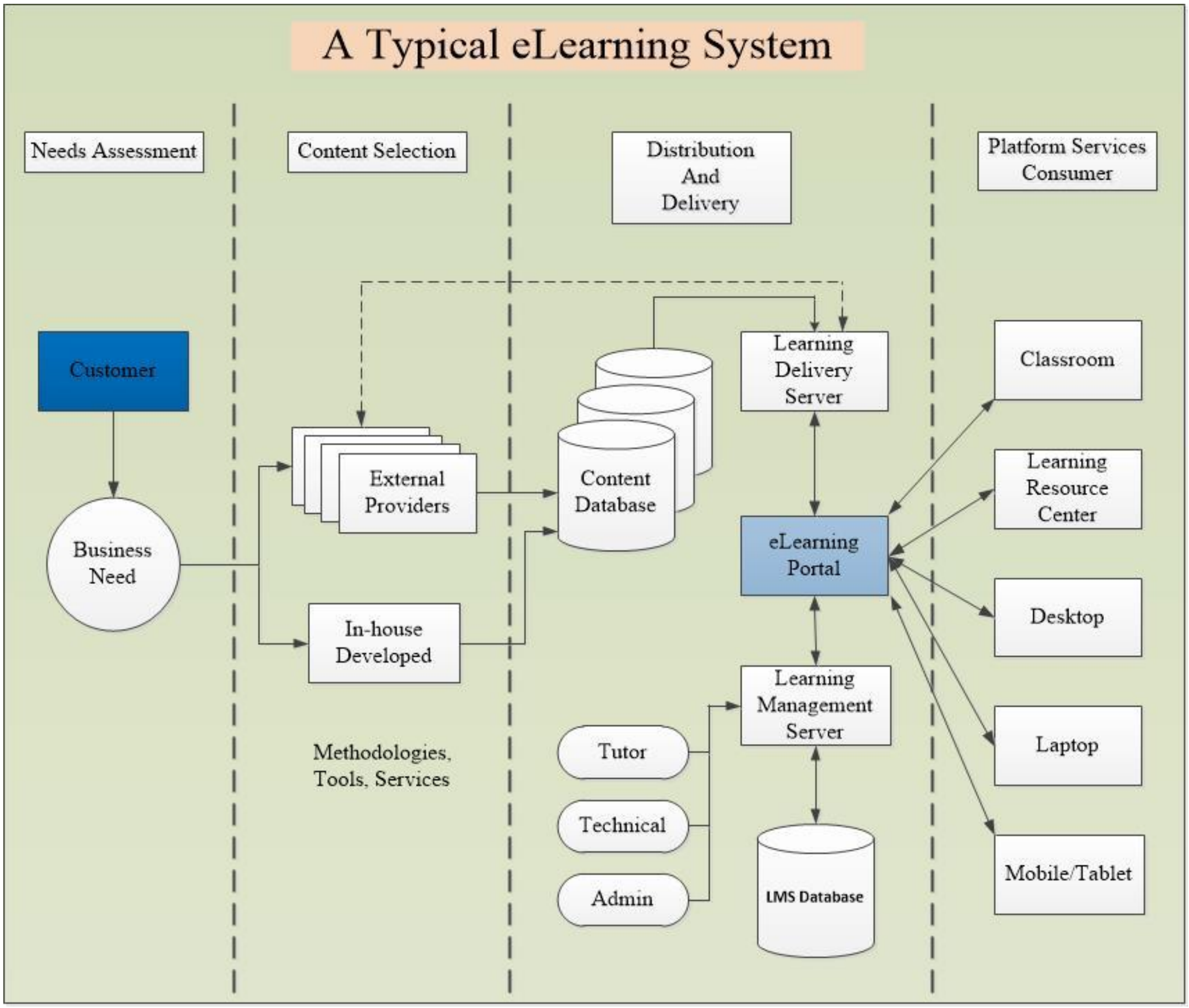


Figure 1: Components of a typical eLearning System

Commercially, there are many eLearning systems available in the market. BlackBoard, Desire2Learn, Canvas, Sakai, eCollege, Saras, etc. are few to name.

6 THE FUTURE OF ELEARNING

eLearning systems are constantly evolving with technology trends in hosting, connectivity, devices and services. Accordingly instructional design models are also improving. A concept that has been a hot topic amongst the eLearning community recently is “big data”. Big data, in terms of the eLearning industry, is the data that is created by learners while they are taking an eLearning course or training module. For example, if a learner is interacting with a training module centered on company policies, his/her progress, assessment results, social sharing, and any other data being produced during the eLearning course is “big data”. The Learning Management System, the eLearning Authoring Tool, social media, multimedia, etc, that have been set by the organization or the eLearning professionals, collect the data.

The term “big data” doesn't only apply to the volume of data itself, but the individual pieces of data that are being collected. These pieces of data can be analyzed to offer organizations or eLearning professionals the opportunity to determine how the learner is acquiring information, at what pace, and to pinpoint any problems that may exist within the eLearning strategy itself.

As a result, educational data mining and learning analytics practices are evolving to enhance eLearning business strategy. Hence, the role of data mining in educational and eLearning systems have become quite critical to increase the system capability and effectiveness of pedagogy being offered.

7 WHAT IS DATA MINING

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). Data mining combines database technologies and statistical methods to enable the exploration and analysis of very large data sets [David M. Levine, David F. Stephan, Kathryn A. Szabat, Page 595].

The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases

Data mining can answer questions that cannot be addressed through simple query and reporting techniques.

8 DATA MINING PROCESS

Data mining process is iterative in nature, and hence the data mining project. The process flow shows that a data mining project does not stop when a particular solution is deployed. The population of data keeps growing in the data warehouse system over the time which can introduce new or more prominent data patterns. Hence, the results of data mining may trigger new business questions or altogether prompt to re-prioritize business questions, which in turn can be used to develop more focused models with different mathematical algorithms [www.docs.oracle.com, http://tinyurl.com/kj7ua97, Data Mining].

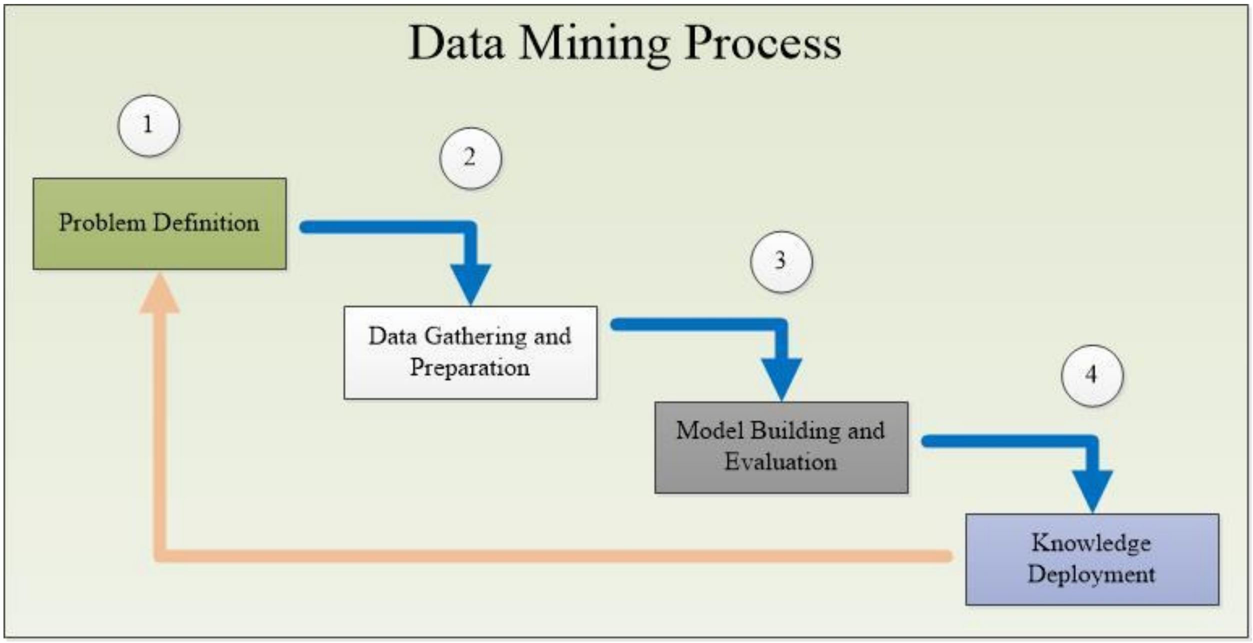


Figure 2: Iterative nature of a data mining process

8.1 Problem Definition

This initial phase of a data mining project focuses on understanding the project objectives and requirements. Once we have specified the project from a business perspective, we can formulate it as a data mining problem and develop a preliminary implementation plan.

For example, our business problem might be: "How can I sell more of my educational products to customers?" We might need to translate this into a data mining problem such as: "Which customers are most likely to purchase the educational products?" A model that predicts who is most likely to purchase the product must be built on data that describes the customers who have purchased the educational products in the past. Before building the model, we must assemble the data that is likely to contain relationships between customers who have purchased the educational products and customers who have not purchased the similar products. Customer attributes might include locality of business, number of years in business, specialty in educational domain, and so on.

8.2 Data Gathering and Preparation

The data understanding phase involves data collection and exploration. As we take a closer look at the data, we can determine how well it addresses the business problem. We might decide to remove some of the data or add additional data. This is also the time to identify data quality problems and to scan for patterns in the data.

The data preparation phase covers all the tasks involved in creating the case table we will use to build the model. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, case, and attribute selection as well as data cleansing and transformation. For example, we might transform a **DATE_OF_BIRTH** column to **AGE**; we might insert the average family income in cases where the **INCOME** column is null.

Additionally we might add new computed attributes in an effort to tease information closer to the surface of the data. For example, rather than using the purchase amount, you might create a new attribute: "Number of Times Students Attempts Tests to Exceed Distinction Grade". Students who assess supplementary learning materials before taking tests may also be related to students who score better than the average students who do not assess supplementary learning materials before the exam.

Thoughtful data preparation can significantly improve the information that can be discovered through data mining.

8.3 Model Building and Evaluation

In this phase, we select and apply various modeling techniques and calibrate the parameters to optimal values. If the algorithm requires data transformations, we will need to step back to the previous phase to implement.

In preliminary model building, it often makes sense to work with a reduced set of data (fewer rows in the case table), since the final case table might contain thousands or millions of cases.

At this stage of the project, it is time to evaluate how well the model satisfies the originally-stated business goal (phase 1). If the model is supposed to predict customers who are likely to purchase an educational product, does it sufficiently differentiate between the two classes? Is there sufficient lift? Are the trade-offs shown in the confusion matrix acceptable? Would the model be improved by adding text data? Should transactional data such as purchases (market-basket data) be included? Should costs associated with false positives or false negatives be incorporated into the model?

8.4 Knowledge Deployment

Knowledge deployment is the use of data mining within a target environment. In the deployment phase, insights and actionable information can be derived from data.

Deployment can involve scoring (the application of models to new data), the extraction of model details (for example the rules of a decision tree or regression algorithm), or the deployment and integration of data mining models within applications, data warehouse infrastructure, or query and reporting tools.

Business Intelligence reporting tools and dashboards can easily display the results of data mining. The choice of such tools is subject to evaluations by the organizations.

9 WHY DATA MINING

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict and classify results patterns based on the historical data. Here are just a few of the most significant reasons for why data mining has the power to enhance eLearning experience and create more effective eLearning environments:

1) What sequence of topics is most effective for a specific student?
2) What are popular courses and programs offered?
3) Which student actions are associated with better learning and higher grades?
4) Which actions indicate satisfaction and engagement?
5) What features of an online learning environment lead to better learning?
6) When are students ready to move on to the next topic?
7) What course should be suggested to students next?
8) When is a student at risk of not completing a course?
9) What grade is a student likely to receive without intervention?
10) Should a student be referred to a counselor for help?
11) What could be the students' dropout rate in the next year?
12) Expand the understanding of eLearning process

Organizations may prioritize their business problems to find solutions to any of the above questions using data mining techniques.

10 HOW DOES THIS PROCESS CAN WORK IN ELEARNING SYSTEMS

Application of Learning Analytics and Educational Data Mining works in a series of steps continually over the period. Learners continue their engagement in the eLearning contexts, and the knowledge derived from data mining is evaluated and applied into the learning services as per the digital content redesign and business strategies. The workflow can be diagrammatically described as follows.

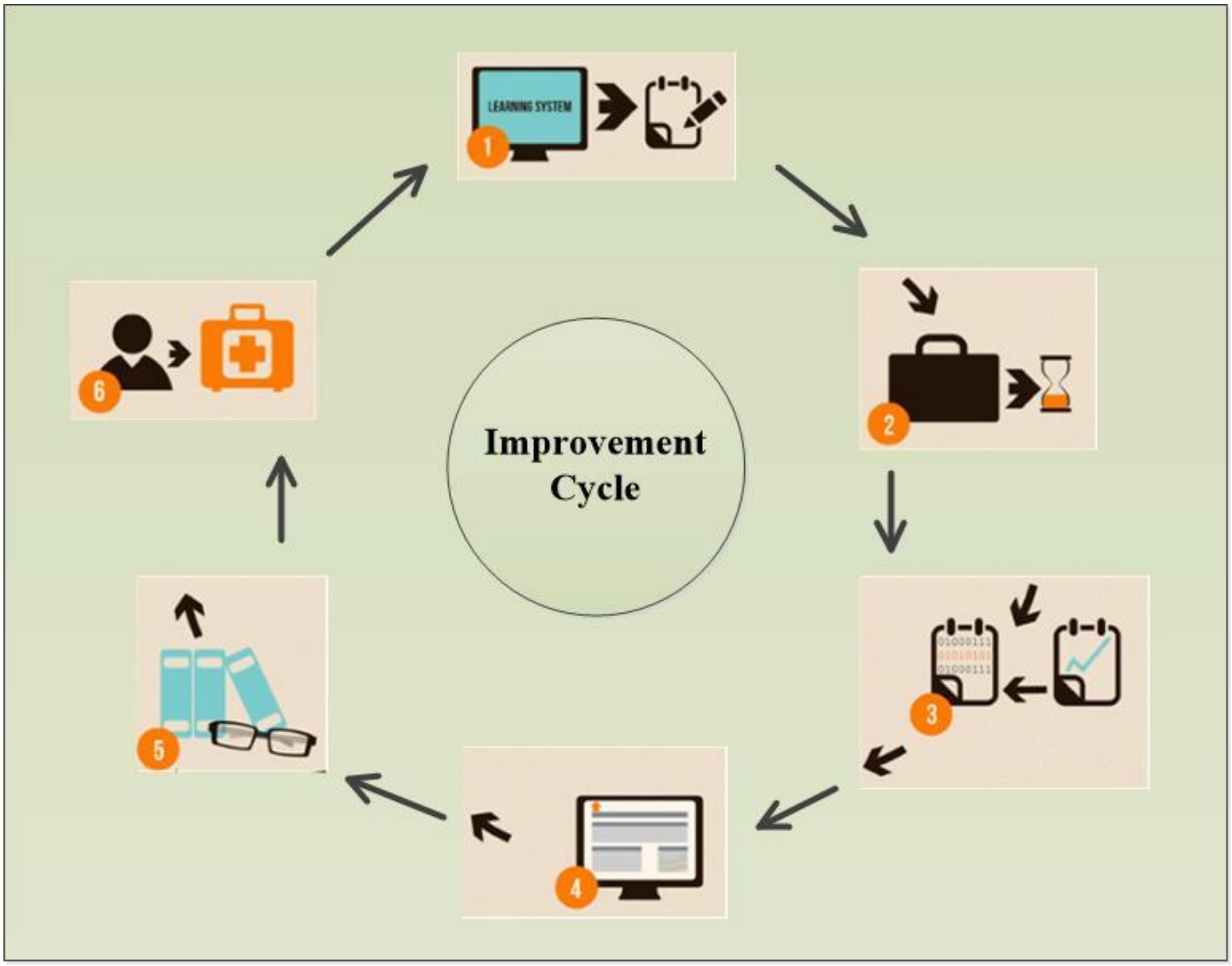


Figure 3: Improvement cycle in eLearning process of adopting data mining discipline

The diagram shows six major steps involved in the process of improving the digital learning and evaluating systems using data mining techniques.

10.1 Step 1 Platform Engagement

Students access their enrolled courses as per their designed learning objectives through digital platforms of eLearning systems like Learning Management System, Test and Assessment System, etc. Students complete their tests and assessments as per schedules prescribed by the course evaluators and instructors. Students achieve certain grades and scores at the end.

10.2 Step 2 Capture Engagement Data

Students activities like login, logout, course navigation, time spent reading learning materials like eBook, time taken to complete tests, etc. are captured and stored in the database by the eLearning systems. Further grades and scores are evaluated and shown to the students, evaluators and instructors.

10.3 Step 3 Data Mining

With passage of time, eLearning systems get to collect lots of such information about students' engagement behaviours, grades, scores, passing rates in certain types of questions, enrollments into various courses and programs, etc. All these data can be used for qualitative and quantitative analysis by data systems experts and extract knowledge of business importance to help content designers, instructors, evaluators and university officials to take further decisions.

10.4 Step 4 Presentation of Data Mining Results

Results of data mining and their interpretations are displayed on the appropriate visuals dashboards like MIS.

10.5 Step 5 Improve Personalized Digital Experience

University officials, product managers, content designers, course authors, business analysts and technical experts intervene to prepare and design better digital experience. Students and instructors receive enhanced digital contents to better suit their interests, skillsets and other needs.

10.6 Step 6 Evaluate and Improvise Platform Services

University officials, product managers, content designers, course authors, business analysts and sales experts can provide further guidance to help instructors and students get better infrastructures and services over the time.

Data mining techniques experts can formulate and identify various correlated data variables to develop data mining models to run through various statistical algorithms. Arriving to a conclusion about any data mining model with certain set of data variables require extensive analysis of engagement data logged by eLearning systems. Various conclusions and inferences can be done during the data analysis process, but product owners are required to focus only those areas of knowledge that can benefit their core business objectives.

Data analysis and decision models need to be evaluated periodically as more data collection happens into the systems over the time. New data patterns and exceptions are likely to generate as engagement data size keeps growing inside the digital systems. Accordingly data mining models, statistical algorithms and priority of knowledge decisions are likely to change for improvements, and hence the performance of the data mining model.

11 DATA MINING MODELS, RESULTS AND INTERPRETATIONS

Data mining model building exercise begins once data source is ready and data available in it. Based on the problem definition, appropriate set of data variables need to be selected for experimenting significance and correlation among them. Data variables are extracted from domain entities of the business. Course, program, users, course/program enrollments, assessments, grades/outcomes, etc. are major form of domain entities found in eLearning systems.

Data mining is an iterative process. Data analysts must explore alternate models that suit the most to the business problem in hand. While exploring for a good model, data analysts may learn a need to make some customizations in data or create new variables (like ratios, differentials, etc.) or even modify the problem statement. With time, data mining model may have to be re-trained with modified parameters or algorithms altogether also as the business problems themselves might change in value to the business.

The paper intends to define a few set of business problems and build data mining models to run through certain data analysis methods.

11.1 BUSINESS SCENARIO

eLearning business has different models of business and hence varying delivery of contents, conducting tests, evaluating grades, tracking progress and reporting. A typical business model is summarized below to understand the data models and their data mining results.

Table 1: Scenario of an eLearning System

A typical eLearning delivery and evaluation system
1. Course contains a list of chapters followed by term and final exams. Each chapter further contains a set of learning objectives, study materials and a practice test. 2. Instructors and teaching aides plan and track students' progress chapter wise. They evaluate progress of students based on a practice test in each chapter before moving on to the next chapter. 3. Students refer to study materials for better understanding before appearing for a practice test in the chapter. Study materials are designed to help students get better understanding of the chapters and courses. Students refer to study materials as per

their convenience online.

4. The time spent on each study material and chapter is termed as 'engagement level'. LMS has capability to track the time duration spent by each student on each study material and chapter.
5. Tests and exams are designed to have specific question types, patterns and their weightages.
6. With chapter wise progress, students finally take mid-term and final exams as per the study plan in the course.
7. The performance of students in each test can be measured chapter wise. Similarly mid-term and final exam help measure students' performance at a course level.
8. Students' grades or outcomes can be later reviewed by instructors and other authorities from different perspectives, say engagement level, content design of study materials, chapters, tests, etc. in a course.

Note: The course, chapter and tests organizations are however dependent on their content design and authoring model.

11.2 Data Mining Techniques

There are a vast range of data mining techniques and associated to analyze data and results. The paper uses some of methods of both descriptive and predictive data mining techniques to demonstrate usefulness in data of educational and eLearning systems.

1.2 Cluster Analysis

The Cluster Analysis is an explorative analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases, i.e., observations, participants, respondents. Cluster analysis is used to identify groups of cases if the grouping is not previously known. Because it is explorative it does make any distinction between dependent and independent variables.

When a cluster analysis is started, data analysts do not know which all attributes can link together to form clusters in what way and how many. Hence, the results of cluster analysis must be interpreted by someone knowledgeable in related business to weigh on attributes contributing to different clusters. Since clustering method segments data into groups not-defined-previously, it is also sometimes referred to as unsupervised learning.

2.2 Regression Analysis

The Regression Analysis uses existing data to forecast what other values will be. It uses predictive models to predict next set of response values based on that are already built or trained in the past. Since the newly calculated or estimated values are compared with the earlier known results, this method of analysis is also called supervised learning.

Both the methods of data mining techniques are employed in the next section for illustration.

11.3 Data Mining Tools

Following tools and services were used to do exercises on data mining to validate selected data mining methods of Regression and Cluster Analysis.

Table 2: Data Mining Tools and Services

S No	Tool Name	Type	Description
1	MS Excel 2010	BI Tool	MS Office Package
2	MS SQL Server Analysis Service 2008	Analysis Service	Analysis service to provide statistical algorithm methods
3	MS Excel Analysis ToolPak	Excel-Addin	Statistical and Analysis Tool for statistical and engineering analysis
4	MS SQL Server (R) Data Mining Excel Add-in	Excel-Addin	Statistical and Analysis Tool for statistical and engineering analysis using Data Mining Structure and complex algorithms

11.4 Business Problems

Table 3: Business Problem 1

Business Problem 1: Predict the student grade based on engagement level inside a course	
Description	Study materials are designed to help students understand the subject and better prepare for the exams. Engagement level of each student in study materials greatly influences the exam performance. Seeing the past trend of engagement level in each chapter, students' grades can be predicted to a large extent.
Data Mining Method	Linear Regression

1.4 Regression Analysis: Data Model

Table 4: Grading results of students scored in a test of a chapter in the course

Student Grades Prediction			
Student Name	**Outcomes of Assessment {n} in a Chapter**	**Time Taken to Submit Assessment (in minutes)**	**Engagement Level (in hours)**
Beau	81	30	10
Katie	80	30	10
Lynnette	75	30	8
Evans	81	29	10
Kent	79	30	8
Jason	87	30	10
Dusko	86	29	9
Peter	85	30	10
jean	79	30	12
Anna	76	30	7

Deanna	76	30	9
Curtis	81	29	13
Frank	82	29	10
Vickie	80	30	6
Stefanie	74	30	6
Chris	70	30	5
MARIE	89	29	11
Alianna	78	30	8
Alfredo	80	30	9
Tyler	77	30	10
Fidias	78	30	8
Brianna	82	29	11
David	77	30	8
Dennis	75	30	7
Patrick	79	29	9
Evelyn	75	30	7
Ashley	74	30	9
Stewart	84	30	10
Larry	68	29	7
Melissa(Lisa)	77	30	8

1.1 Results and Interpretations

Regression Result 1

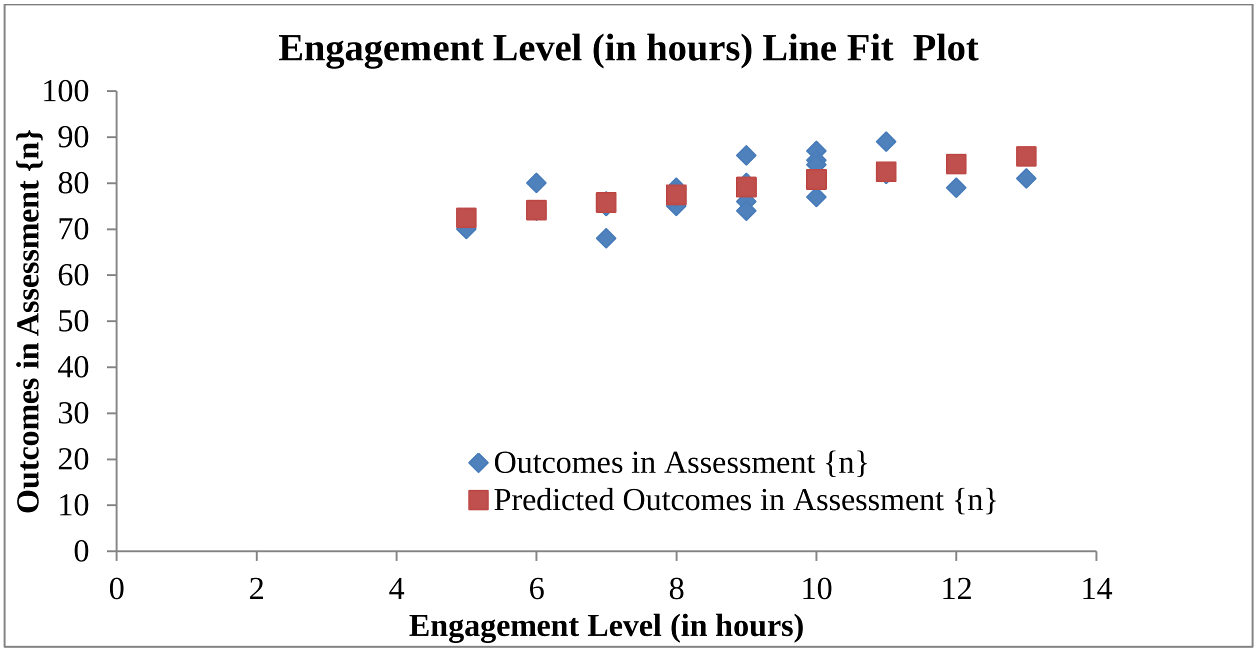


Figure 4: Correlation between engagement level and outcomes

1.1 Interpretation of the Line Fit Plot

i. Outcomes are aligned in the line direction of engagement level. Meaning, engagement level and outcomes are correlated [Barry Render, Ralph M. Stair, Jr Michael E. Hanna, T. N. Badri, Page 122].

ii. With increase in engagement level, outcomes values also increase positively.

Table 5: Business result of correlation between engagement level and outcomes

Business Advantages and Implications for Practitioners
i. Fit line of engagement level through the grades distribution can tell how effective has been study materials benefitting students in tests of each chapter of the course.

Regression Result 2

Table 6: Regression output generated in MS Excel BI Tool

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.646536128							
R Square	**0.418008965**							
Adjusted R Square	0.397223571							
Standard Error	3.642052399							
Observations	**30**							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	266.7594	266.7594	20.1107	**0.00011**			
Residual	28	371.4073	13.2645					
Total	29	638.1667						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	64.12132	3.34735	19.15585	0.00000	57.26458	70.97805	57.26458	70.97805
Engagement Level (in hours)	1.66551	0.37139	4.48450	**0.00011**	0.90475	2.42628	0.90475	2.42628

1.2 Interpretation of Regression Output

i. The coefficient of determination (R^2) is on higher side 0.4180, though value closer to 1 is desirable.

ii. R^2 of 0.4180 means 41.80% of variability in grades distribution can be explained by the above data model.

iii. The overall model is helpful because the ‘Significance F’ probability is low 0.00011% (much less than 5%) [Barry Render, Ralph M. Stair, Jr Michael E. Hanna, T. N. Badri, Page 123].

iv. Attribute ‘Engagement Level’ is helpful because the ‘p-value’ is low 0.00011% (much less than 5%).

1.3 Point Prediction

Table 7: Point prediction output generated in MS Excel BI Tool

Grade Prediction against new Engagement Level Data							
					Approx. 95% of PI		
			Approx. St. Error	**Margin of**	**Lower**	**Upper**	**Interval**
Engagement Level (in hours)	**Point Prediction**	**t-value**	**Prediction**	**Error**	**Bound**	**Bound**	**Width**
10	80.78	2.0484	4.0063	8.2064	72.5700	88.9829	16.4129
9	79.11	2.0484	4.0063	8.2064	70.9045	87.3174	16.4129
8	77.45	2.0484	4.0063	8.2064	69.2390	85.6519	16.4129
11	82.44	2.0484	4.0063	8.2064	74.2355	90.6484	16.4129
7	75.78	2.0452	4.0063	8.1937	67.5862	83.9736	16.3874

The above table shows the point prediction values calculated using regression output in Table B for new set of engagement level data in Table C.

New engagement levels of 10, 9, 8, 11 and 11 predict grades of 80.78, 79.11, 77.45, 82.44 and 75.78 respectively.

Hence, comparing the above point prediction results in Table C as against know values of Grades vs Engagement Level in trained model data Table A, Table C prediction can be said to be correct. That is, the trained model is able to predict future unknown values, *response*.

Table 8: Business result of point prediction

Business Advantages and Implications for Practitioners
i. With the trained data, grades in chapters can be predicted measuring the engagement level of students enrolled into the course. ii. Feedbacks or counseling can be given to those students that are likely to fare poor in the exam seeing their engagement levels below the desired hours.

Table 9: Business Problem 2

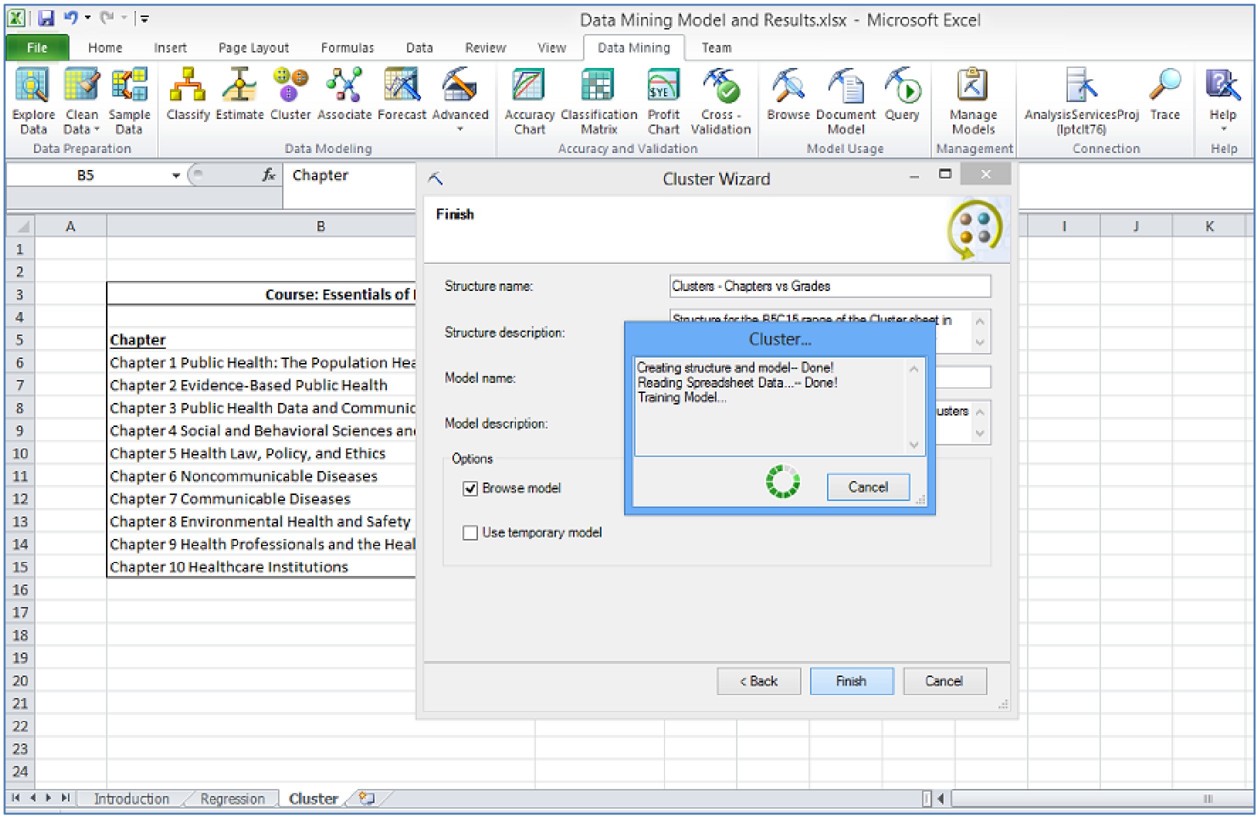
Business Problem 2: Benchmark student grades for comparison and review	
Description	Average grades in each chapter of a course can help identify the potential grades by each student seeing his engagement level in the chapter. Results can also help to evaluate if study materials are effective enough to attract students.
Data Mining Method	Cluster Analysis

2.4 Cluster Analysis: Data Model

Table 10: Average grades chapter wise

Course: Essentials of Public Health	
Chapter Titles	**Average Grades %**
Chapter 1 Public Health: The Population Health Approach	76
Chapter 2 Evidence-Based Public Health	74
Chapter 3 Public Health Data and Communications	79
Chapter 4 Social and Behavioral Sciences and Public Health	75
Chapter 5 Health Law, Policy, and Ethics	67
Chapter 6 Non-communicable Diseases	69
Chapter 7 Communicable Diseases	71
Chapter 8 Environmental Health and Safety	82
Chapter 9 Health Professionals and the Health Workforce	74
Chapter 10 Healthcare Institutions	81

2.1 Results and Interpretations



B5: Chapter

	A	B
1		
2		
3		Course: Essentials of
4		
5		Chapter
6		Chapter 1 Public Health: The Population Hea
7		Chapter 2 Evidence-Based Public Health
8		Chapter 3 Public Health Data and Communic
9		Chapter 4 Social and Behavioral Sciences an
10		Chapter 5 Health Law, Policy, and Ethics
11		Chapter 6 Noncommunicable Diseases
12		Chapter 7 Communicable Diseases
13		Chapter 8 Environmental Health and Safety
14		Chapter 9 Health Professionals and the Heal
15		Chapter 10 Healthcare Institutions

Cluster Wizard

Finish

Structure name: Clusters - Chapters vs Grades

Structure description: Structure for the B5C15 range of the Cluster sheet in

Model name:

Model description: usters

Options

☑ Browse model

☐ Use temporary model

Cluster...

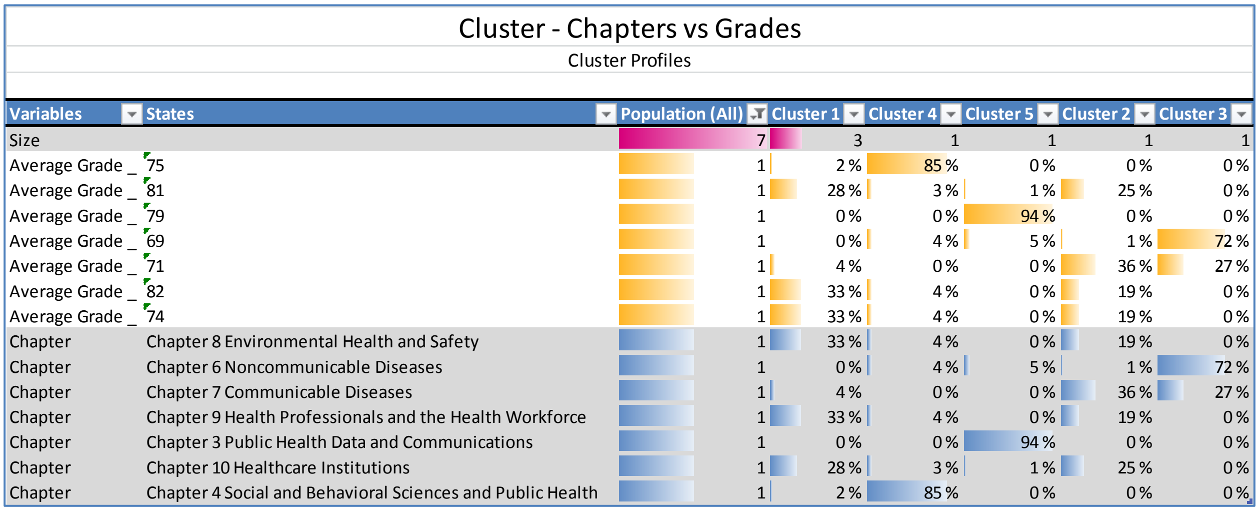
Creating structure and model-- Done!
Reading Spreadsheet Data...-- Done!
Training Model...

Cancel

< Back | Finish | Cancel

Figure 5: Creating clusters in MS Excel 2010 and MS Analysis Services 2008 R2

Cluster Result 1


Cluster - Chapters vs Grades

Cluster Profiles

Variables	States	Population (All)	Cluster 1	Cluster 4	Cluster 5	Cluster 2	Cluster 3
Size		7	3	1	1	1	1
Average Grade _	75	1	2 %	85 %	0 %	0 %	0 %
Average Grade _	81	1	28 %	3 %	1 %	25 %	0 %
Average Grade _	79	1	0 %	0 %	94 %	0 %	0 %
Average Grade _	69	1	0 %	4 %	5 %	1 %	72 %
Average Grade _	71	1	4 %	0 %	0 %	36 %	27 %
Average Grade _	82	1	33 %	4 %	0 %	19 %	0 %
Average Grade _	74	1	33 %	4 %	0 %	19 %	0 %
Chapter	Chapter 8 Environmental Health and Safety	1	33 %	4 %	0 %	19 %	0 %
Chapter	Chapter 6 Noncommunicable Diseases	1	0 %	4 %	5 %	1 %	72 %
Chapter	Chapter 7 Communicable Diseases	1	4 %	0 %	0 %	36 %	27 %
Chapter	Chapter 9 Health Professionals and the Health Workforce	1	33 %	4 %	0 %	19 %	0 %
Chapter	Chapter 3 Public Health Data and Communications	1	0 %	0 %	94 %	0 %	0 %
Chapter	Chapter 10 Healthcare Institutions	1	28 %	3 %	1 %	25 %	0 %
Chapter	Chapter 4 Social and Behavioral Sciences and Public Health	1	2 %	85 %	0 %	0 %	0 %

Figure 6: Cluster profiles

1.1 Cluster Interpretation

i. Cluster 1 has higher grades distributions of 28%, 33% and 33% in chapters 8, 9 and 10 respectively.

ii. Cluster 2 has higher grades distributions of 25%, 36%, 19% and 19% population in chapters 10, 7, 8 and 9 respectively.

iii. Cluster 3 has higher grades distributions of 72% and 27% population in chapters 6 and 7 respectively.

iv. Cluster 4 has a higher grade distribution of 85% population in only chapter 4.

v. Likewise, cluster 5 has a higher grade distribution of 94% population in chapter 3.

vi. However, grades of tests in chapters 1, 2 and 5 have been less impressive compared to other chapters in the same course. Instructors and other evaluators should analyze the causes of it.

Cluster Result 2

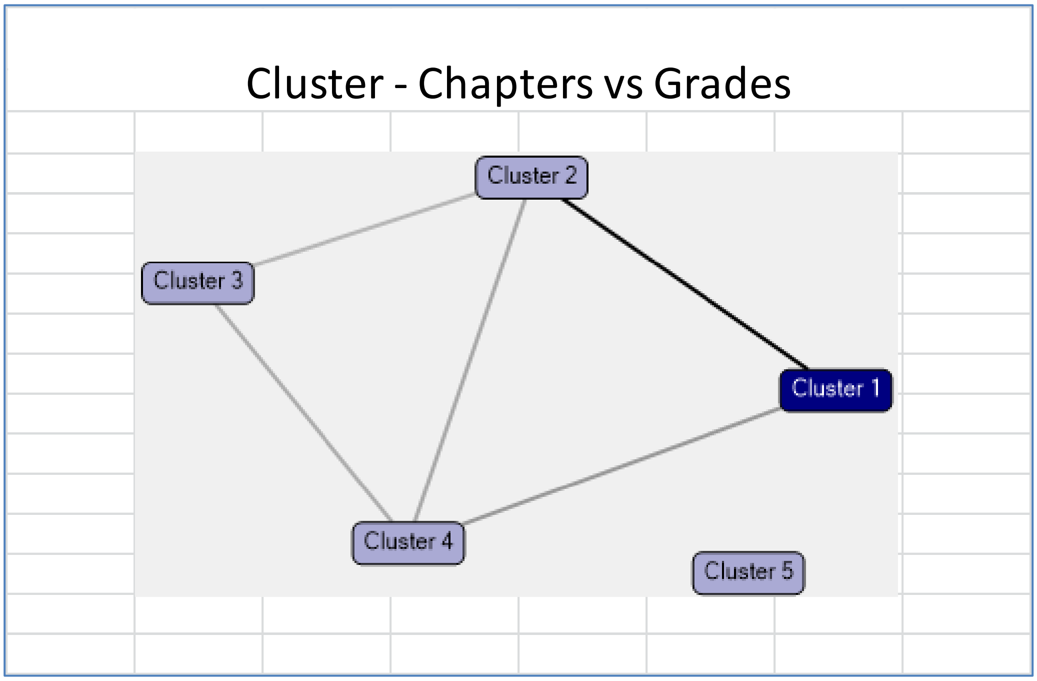


Figure 7: Cluster diagram showing strong link between Cluster 1 and Cluster 2

1.2 Interpretation of Cluster Links

i. Clusters 1 and 2 have higher number of grades distribution patterns across multiple chapters among students.

ii. Cluster 5 seems to have weak link in relation to other clusters. **Cluster 5 needs immediate attention from instructors and other evaluators.**

Table 11: Business result of cluster analysis

Business Advantages and Implications for Practitioners
i. Each cluster can help identify chapter wise class average grades distribution. ii. Density of each cluster can tell the effectiveness of the particular chapter(s) content and delivery. iii. Clusters can help identify semester or year wise comparison of the course, further it can help identify any unusual patterns in class average grades so that further drilling can be done to know other causes of class performance.

The above case study shows how data mining techniques can help confirm certain empirical observations and find new, subtle patterns and occasional breakthrough insights of business importance. Organizations can miss opportunities to benefit and improvise on time if such hidden knowledge is left un-traversed only because of lacking learning analytics discipline in the business strategy and decision making activities.

12 DATA MINING: WHAT IT CANNOT DO

Data Mining is not an automated tool like thing that can continuously watch the data and report about an interesting pattern of knowledge. It can uncover hidden knowledge and patterns from the pile of data, but it cannot validate business value of the discovered knowledge at its own- business experts need to intervene and provide guidance time to time to data analysts and data scientists.

Data Mining does not replace skilled business analysts or managers, but rather it gives them a powerful new tool to improve the job they are doing. Data Mining utilizes the computing power of computers and return results of statistical methods in much lesser time; otherwise what could have been tasks of years by manual intervention.

13 COSTS AND CHALLENGES OF DATA MINING DISCIPLINE

There are certain costs and challenges involved in developing infrastructures to be ready enough to practice data mining techniques in the historical data stored by eLearning systems. Data analytics being the evolving subject area in educational systems, identification of data models that can apply seamlessly across all eLearning systems is still a major challenge as all platform services do not operate or specialize in common set of services- their specialization and services to end users vary largely and they leverage platform services by integrating with multiple other platforms to provide a rich digital experience. The following points can be key costs and challenges any eLearning systems that may wish to leverage data mining techniques into their business.

13.1 LACK OF BUSINESS QUESTIONS

Most organizations do not have business questions though they may be rich in business data. They believe data and infrastructure costs are limitations to adopting learning analytics in their organizations. However, they should understand having business questions first is the key driver to start with learning analytics in their business.

13.2 COSTS TO STORE LOGGED DATA

Recording and capturing learners' engagements and their activities in the eLearning systems applications require intelligent logging and instrumentation of key data variables in the database that can be further associated to learners' sessions in the application. Application code needs to support efficient logging of data for each user's session. The other cost factor includes database server maintenance, database backup, data extraction, data migration and at times data sharing across platforms. eLearning systems that have no support of such logging and instrumentation in their application code may have to refactor code in application layers, which can be a cost burden if they wish to have analytics capability in their application data.

13.3 COSTS TO ADOPT EMERGING STANDARDS OF ENGAGEMENT DATA AND COMMUNICATION

As the saying goes- software makes services better, and services make software better. The notion very much applies to SaaS model of software development and its application. Application code no longer runs only on desktop. The platform varies widely from desktop to web to mobile, tablets and other held devices. Accordingly data generated from all these various medium has to be of some notational standards in order to be light weight, fast, veracious and widely accepted across discrete systems.

Emergence of eLearning has led to development of a new standard of engagement data format called Tin Can API, aka xAPI. It enables the collection of data in a typical JSON string format about a wide range of learning experiences a person goes through. It relies on a Learning Record Store (LRS), and it overcomes the majority of limitations of SCORM, which was the previous standard. Engagement data formatted and stored into xAPI standards are highly efficient and cost effective in terms of data communication, persistence, interoperability and mapping against domain entities inside the database. But many older eLearning platforms may not be ready with this technology adoption. However, some are developing a plugin connector to capture engagement data into xAPI formats and persisting separately into the data warehouse without altering the existing architecture of core platforms. The cost is obvious there.

13.4 COSTS TO HIRE DATA SYSTEMS PROFESSIONALS

Experts like data scientists, data analyst and domain experts are required to understand and extract knowledge from data. Only extracting knowledge may not be sufficient when it cannot be correlated to business objectives. It is not only hiring of such professionals becomes difficult, but also retaining them in the long run who can be aligned with organizations growth plan.

13.5 DATA SYSTEMS INTEROPERABILITY

Many eLearning systems run typically in their niche areas and capabilities to meet market needs. Some systems may specialize in content designing, some in tests and assessments, some in specific type tests only, some in specific subjects, etc. But when the customer needs grow for a range of services under one roof, the platform owners integrate their services for one another to meet customer demand. But all such integrated platforms most likely may not have interoperable data formats, entities life cycle, etc. Hence, data systems interoperability can become a major challenge in designing data warehouse and data mining models.

13.6 CREATING ONE SIMPLIFIED PICTURE FROM DATA CAN BE DIFFICULT

Learners' behaviour is a subjective context that can be attributed to various factors of their educational, professional, demographic, etc. details. The correlation effects of some data variables for certain set of sample population may not be equally true to various other samples. eLearning systems, being digital platforms equally accessible to all, witness customers coming from too diversified demography. Hence, the complexity of knowledge inferences may grow with the size and volume of learners' engagement data. Data analyst, domain experts and business intelligence experts may have to experiment with multiple data mining models validating, training, testing and applying into reporting tools.

13.7 PRIVACY AND ETHICS NEED TO BE UPHELD

Data privacy of an individual is of utmost importance as an application security. Schools and universities have records of their students and teachers that are more than personal like family, demographic, interests, hobbies, culture, etc. Data mining requires detailed analysis and direct exposure to users' data of all nature, not limited to grades and scores only. There would be a risk of racial discrimination or risk of demographic profiling or illegitimate kind of conclusions about users. Hence, the issue of data privacy and ethical practices need to be addressed by organizations developing analytics capability into their systems.

13.8 CHOOSING DATA MINING MODEL AND APPROPRIATE ALGORITHMS

Selection of data mining model and appropriate algorithms require intensive exercise of data analysis, data variables selection, data cleansing, finding data exceptions and building right algorithms to apply appropriate statistical algorithms. Accuracy of data, defining errors threshold, training data model, testing results, etc. are done in iterations over long period before deriving a business knowledge of importance.

The selection of business intelligence enterprise tools and their infrastructure cost also have impact on the selection of data mining model.

14 CONCLUSION AND FUTURE WORK

Data mining techniques, the core art of Learning Analytics, offer great promises in helping organizations discover patterns and insights hidden in their pile of data that can have potential to benefit business and customers in many ways. However, data mining techniques and tools employed by data analysts and data scientists need to be guided by business experts who can find relevance in outcomes of data analysis.

Learning Analytics is the next big trend to define the methodology of online learning, and hence the size of business, its edge and competency in the eLearning market.

Learning Analytics is a promising field for both business and business professionals. Various data models and other methods of data mining techniques should be explored to understand the usefulness of this discipline.

15 BIBLIOGRAPHY

David M. Levine, David F. Stephan, Kathryn A. Szabat, Statistics for Managers using Microsoft Excel Seventh Edition, PHI

Barry Render, Ralph M. Stair, Jr Michael E. Hanna, T. N. Badri, Quantitative Analysis for Management Eleventh Edition, Pearson

Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crows Corporation, www.twocrows.com

Data Mining Overview at www.docs.oracle.com, http://tinyurl.com/kj7ua97

The Research Bulletin of Jordan ACM, Vol. II (III), Improving Student's Performance Using Data Clustering and Neural Networks in Foreign Language Based Higher Education, http://ijj.acm.org/volumes/volume2/no3/ijjvol2no3p1.pdf

Alaa El-Halees, Mining Students Data to Analyze Learning Behavior: A Case Study, Department of Computer Science, Islamic University of Gaza, Palestine, https://uqu.edu.sa/files2/tiny_mce/plugins/filemanager/files/30/papers/f158.pdf

16 REFERENCES

How Can Educational Data Mining and Learning Analytics Enhance Education Infographic, http://elearninginfographics.com/how-can-educational-data-mining-and-learning-analytics-enhance-education-infographic/

Training & Learning Architecture (TLA): LearningRecordStore, http://www.adlnet.gov/tla/lrs/

Data Mining Concepts, http://docs.oracle.com

Machine Learning Blog, http://blogs.technet.com/b/machinelearning/

Prediction using Linear Regression, step-by-step, https://www.youtube.com/watch?v=nFj7nAeGlLk

Prediction using Multiple Regression, step-by-step, https://www.youtube.com/watch?v=E73AJ73-S6g

ProfTDub Series, https://www.youtube.com/results?search_query=profTDub+video+regression

Regression Output Analysis, http://www.statisticshowto.com/excel-regression-analysis-output-explained/

17 NOTES