



LOVELY
PROFESSIONAL
UNIVERSITY

Transforming Education Transforming India

BEHAVIOURAL ANALYSIS OF E – LEARNERS
(Using Web Mining)

Submitted to

LOVELY PROFESSIONAL UNIVERSITY

In partial fulfillment of the requirements for the award of degree of

MASTER OF COMPUTER APPLICATION (HONOURS)

Submitted by

Sushmita Sinha - (11410153)

Supervised By

Ms. Navpreet Kaur
(Assistant Professor, SCA)

LOVELY FACULTY OF TECHNOLOGY & SCIENCES
LOVELY PROFESSIONAL UNIVERSITY
PUNJAB
[APRIL 2015]

DECLARATION

I hereby declare that the Paper writing entitled, **Behavioural Analysis of E-Learners** submitted for the Master of computer application (Hons) Degree is entirely our original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Sushmita Sinha (11410153)

CERTIFICATE

This is to certify that Sushmita Sinha pursuing MCA (Hons.) paper writing titled, **“Behavioural Analysis of E-Learners”** under my guidance and supervision. To the best of my knowledge, the present work is the result of their original investigation and study. No part of the paper writing has ever been submitted for any other degree or diploma. The paper writing is fit for the submission and the partial fulfilment of the conditions for the award of MCA (Hons.).

We further declared that we or any other person has not previously submitted this report to any other institution/university for any other degree/ diploma or any other person.

Dated :

Signature of Supervisor
Ms. Navpreet Kaur
(Assistant Professor, SCA)

TABLE OF CONTENT

TITLE	PAGE NO
1. Introduction	05
1.1 Background	06
1.1.1 Traditional vs E-Learning	06
1.1.2 Synchronous vs Asynchronous e-Learning	07
1.2 Introduction to Data mining	08
1.3 Introduction to Cluster Analysis	10
2 Literature Survey	11
3 Present Work	14
3.1 Problem Formulation	14
3.2 Research Methodology	15
4 Result and Discussion	18
5 Conclusion and Future Scope	27
5.1 Conclusion	27
5.2 Future Scope	27
6 References	28
6.1 Article	28
6.2 Papers	28

BEHAVIORAL ANALYSIS OF E – LEARNERS

(Using Web Mining)

Sushmita Sinha, Rahul Hajong
School of Computer Application
Lovely Professional University, Punjab, India

Chapter 1

INTRODUCTION

In today's era, the popularity of getting digitalized has become the significant trend, hence the emergence of web-based education or E-learning has become the growing demand of the people. E-learning has given the world a totally different platform from old educational concept. E-learning provides an environment which is completely learner oriented which delivers all the desired digital content for the student as well as teachers. This platform provides more realistic learning society which is typically based on life-long learning. This learning procedure will need tools and techniques that would present the knowledge, make an interactive environment and also would be sufficient enough to share it with others. With this regard, the e-learning concept has become an important tool that can support the learning system and fulfill the desired goals. These technologies will create a dynamic learning context that would facilitate the learning anytime and anywhere. The user (mainly students and teachers) will be able to grab the online learning contents through PC's, mobile phones and other handheld devices.

E-learning has many of its own definitions that give you a complete understanding of this approach. It includes:

- Online learning/education
- Distance education/learning
- Technology based training
- Web based learning/training
- Computer based training/ learning from a CD-ROM

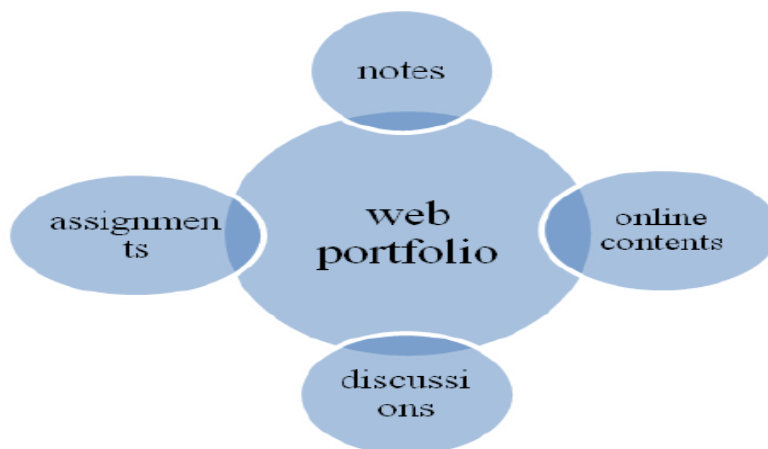


Figure 1.1 E-learning systems

1.1 Background

We use a huge variety of electronic media that can support and deliver the E-learning environment to the user, which is the ideal complement to a traditional education or training program. Earlier, the traditional classrooms had the usual teacher who used to communicate more than students and teacher taught the chapters as per the instruction plans prepared earlier by the study program and existing curriculum. But unlike the traditional learning, e-learning system is based on the technology that is computer based or web based learning. The technology would consist of a virtual classroom and the interaction took place through digital media. The student determines their subject interest and their study will be further based on various sources of information like web data banks and online experts' help. The approach focuses on students' learning "how" rather than "what"; the student is made to do a research study on learning topics that merge searching for and collecting information from web data banks and authorities on the communications network; the approach is marked better when the learning is connected to the real world scenario because of which the subject content is enriched and also it holds the materials in different formats. It is certain that the involvement of computers in the field of education curriculum has marked to be the most significant developments for all the users, being Teacher or Student. For instance, e-learning also enables a forum where each student can know other student's view on same query. And the gracious part of e-learning is that it is accessible anytime and from anywhere, helping in time management as well as place management problems.

1.1.1 Traditional vs. E-learning

Initially, the traditional learning were criticized that there would be no actual interaction between the instructor and the learner and the learning system would be less effective. Moreover, it would not give

The opportunity for shy students to emerge with their problems in-front of huge amount of pupils. But E-learning gives this great advantage for such type of student to get resolved with all types of doubt by private conversations like email, chats, far from physical classroom. It would not only keep the record of the conversation made by the learner and instructor but would also motivate the students and they will not hesitate in asking the queries. There are a lot of benefits of e-learning system which includes:

Flexibility: E-learning is anytime anywhere learning system so it is flexible to learn while convenient.

Accessibility: Students can attend classes and access content which is not locally available.

Global Boundaries: Students from different background or geographical location can participate as well as can share their ideas. However, online learning provides students more freedom and less pressure, there is a possibility of inefficient learning or their attitude becomes negative towards learning. It will be the case when instructor do not recognize the quality of learning attained by the students/learners immediately, but if teachers analyze their students' learning situations time to time their learning will not be ineffective. The other disadvantage that hinders is the familiarity of the communication technology that are being used in-order to achieve the goals otherwise it would simply be a waste of time. This study will analyze the students portfolios, grading system will be there. The performance of the students will be identifying on the bases of their test marks, assignment marks, examination record and online record. If students do not focuses on the learning and their performance is not efficient then some warning messages will be send by teachers to improve their learning records.

1.1.2 Synchronous vs. Asynchronous e-Learning

E-learning can be categorized into two, **Synchronous e-Learning** where training is done in real time which is facilitated by instructor. Such type of learning is commonly done over the Internet by through different communication tools. The learners can log in at any particular time and then communicates with the instructor as well as other students and share their views. Synchronous e-Learning also can be done by telephonic conversation, video calling, or two-way live television broadcasting between instructors and learner at any location and any instance of time.

Asynchronous e-Learning comprises usage of CD, DVD, Intranet or Internet based which focuses on the type of work done by student and their self-study program. It includes communication between learner and instructor via online bulletin boards, various discussion groups, and e-mail system. Programs are also self-contained and have links and references so that students can easily solve their queries without any help of instructor. Asynchronous e-Learning gives opportunity for students to learn "anywhere", "anytime" as long as they have compatible equipment with them.

Successful online learning depends on various factors. The most important factor is course quality. Moreover students' effort as well as their characteristics is also part of successful online learning. Student psychological features, attitude whether its positive or negative, and other general abilities comes under these characteristics. For example, the meritorious students who achieve higher positions in old learning methodology are recorded likely to be successful in online courses than other under-performed students. We have taken Grade point average (GPA) as a key to bring seriousness in these online courses. Only the general ability learning may not take you to a successful and enjoyable learning approach. The student characteristics are equally important to be checked. The given list describes the possible parameters that is required to remain successful in online courses.

Self-Motivation: Someone's' ability and interest is the most important for learning on its own.

Independence: Ability to learn independently with least possible structure.

Good Thinking Skills: Ability to think in better way which helps to succeed while learning, For example: Analysis skills and evaluation skills.

Good Time Manager: Establish goals, develop a schedule and complete work before deadline results in successful online learning.

Good Problem Solver: Includes the ability that can troubleshoot and resolve the problems quickly and easily.

Basic Computer Knowledge: The most important trait for successful online course is users must have computer knowledge. Basic computer hardware and software knowledge must be there for online courses.

Some attributes like self-motivation and independence are pre-acquired and others are to be developed and sharpened. A student need to possess high order thinking skills. The technique of Time Management and Problem Solving Strategies can even be induced and the Computer Knowledge can be learned easily within few span of time. Information related for online learning skills can easily acquire from various E-learning sites.

In addition, E-learning system provides same instruction plan to all the students who have opted for the similar course but the students' individuality and capability is different for all which may result to difference in performance of all. Collectively every students' performance can be checked through their grades. Therefore, this study does the cluster analysis of students' data based upon their grades using different data mining tools and techniques.

1.2 Introduction to Data mining

Nowadays, in big organizations there is huge amount of data. The information behind the data is usually unknown. To extract the useful information or hidden predictive information from this previously unknown data we use data mining done over large databases.

The origination of Data Mining has come from Knowledge Discovery Data (KDD) process. It is the analysis step of knowledge discovery from databases. The main purpose of KDD process is to retrieve the useful knowledge or information from the data considering the large databases. It is to be noted that the terminology like *knowledge discovery* and *data mining* have a different approach. The complete process of KDD is to discover and retrieve the useful information from the data. It also includes the evaluation as well as manipulation of the pattern in which useful knowledge has been extracted in-order to extract some sequential information. The KDD also includes the pre-processing of data i.e. data cleaning (which filters noisy and outlier detection and also handles the missing value issues) before the data mining step.

Whereas data mining refers to applying the algorithms to extract patterns from data without any additional steps which include transformation or reduction of the KDD process.

KDD involves following steps:

1. The first step is development and learning of

- The domain of application
- The appropriate prior knowledge
- The demand at the user-end.

2. Create a target data set:

Select a data-set (which focuses on a subset of variables, or data samples) in which discovery of knowledge is to be performed.

3. Data cleaning and preprocessing.

- Removal of noisy and incomplete information from data.
- Handle missing data fields.

4. Data reduction and projection.

- Find useful features and patterns to represent the data depending on the goal of the task.
- Use dimension reduction methods to reduce the number of variables under consideration or to find invariant representations for the data.

5. Choose the data mining task.

Decide which task of data mining is to be used.

Following are the main tasks of data mining:

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

6. Select the data mining algorithm.

- Select methods that are to be used for searching the patterns in data.
- Decide models and attributes for patterns.
- Matching a data mining method with the fulfillment of the KDD process.

7. Data mining.

- Searching for patterns in the way we can represent data such representations as classification rules or trees, regression, clustering, and so forth.

8. Interpretation of mined patterns.

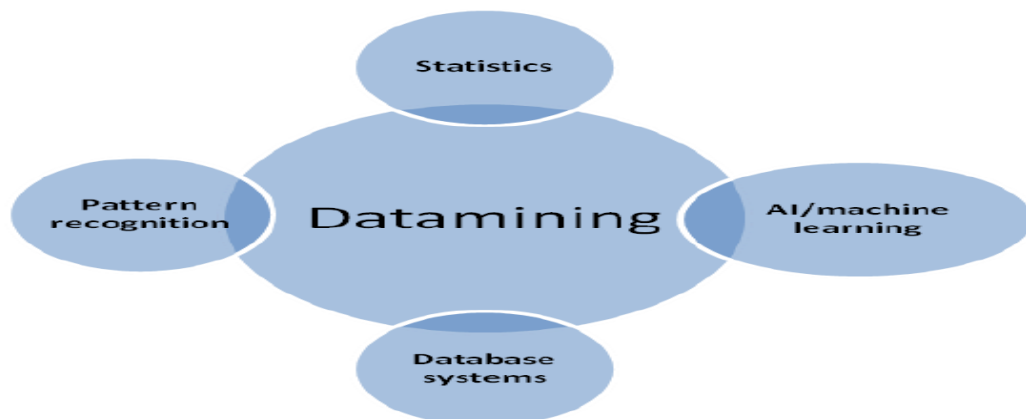
9. Discover knowledge.

1.2.1 Data mining origins:

Data mining gives the concept from different fields such as machine learning/artificial intelligence, statistics, and pattern recognition and database systems.

Challenges of data mining:

1. Scalability
2. Dimensionality
3. Quality of data
4. Data distribution
5. Security and privacy
6. Streaming data
7. Complex data
8. Heterogeneous data



1.3 Introduction to Cluster Analysis

Cluster Analysis is the statistical methods of distributing the information in such a way that the homogeneous data are classified in a single section. This would lead to the collection of similar data groups basically known as clusters. In data mining, clustering is the tool that is used to analyze the data to solve the various problems related to the classification. It would result to classification of set of observations into two or more groups since the degree of association is strong between the same group and cluster and weak between members of different group. Different data mining clustering algorithms are available like k-means clustering, fuzzy c-means clustering and hierarchical clustering algorithm.

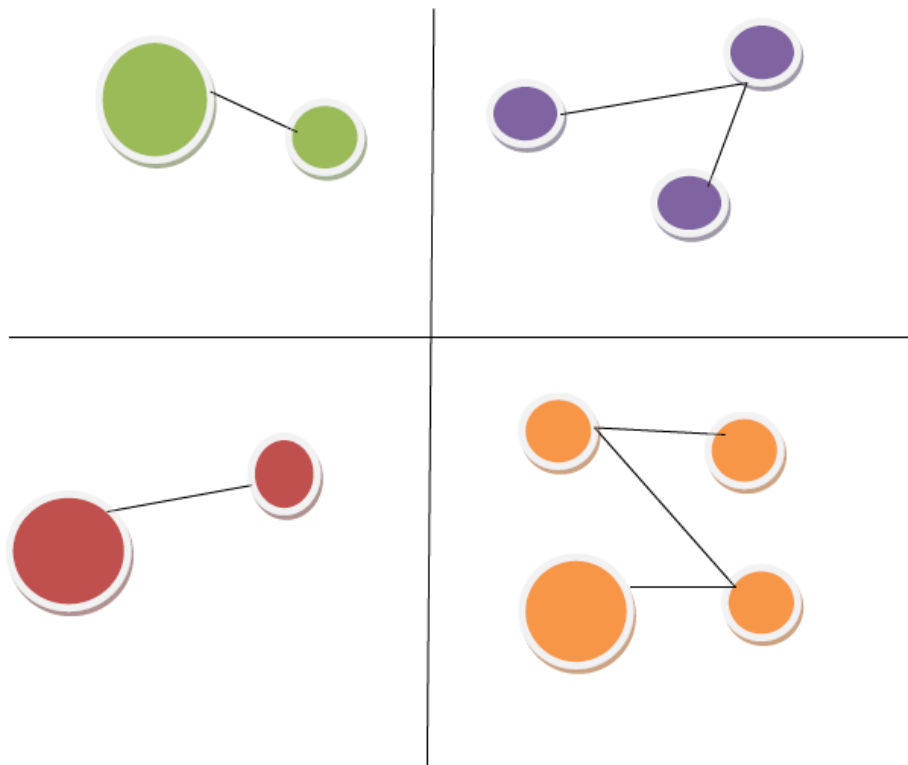


Figure 1.4 Cluster Analyses

1.4 Study of correlation between students' learning behavior and their achievements

To understand the correlation between the learning behavior of the learner and their accomplishments of learning portfolio analysis system that is to be built in-order to enable the instructors to control over the students' personal learning situations. We can also check the procedure of immediate guidance provided by the instructor which can promote and improve the guiding methodology done by them. Learning portfolios helps the instructor and learner to bring the required change in the learning styles.

This study analyzes the behavior of students and their achievements by using various data mining tools and techniques. The students' portfolios help teachers to analyze the students' learning performance and achievements. Using cluster analysis of behavior of student's teachers can change their learning style as well as course contents for students according to their behavior. Cluster analysis categorizes the students according to their grades. The study firstly classify the student's data according to their performance activity whether it's high or low. The goal is that the students with similar level of performance grouped in one cluster and dissimilar grouped in another.

In Literature Survey, I have gone through different works implemented on web-mining and done the analysis over the technologies and methodologies used for web-mining. Referring to those technologies, I have further checked the improvement scope that could be made in this field of web technologies.

1. By S.S. Dhenakaran, S. Yasodha / (IJCSIT) International Journal of Computer Science and Information Technologies.

Semantic Web Mining - A Critical Review

This paper features with the recent issues occurring in World Wide Web, since it contains huge bulk of unstructured data which are only understood by humans. The main focus of semantic is to resolve this problem by providing machine interpretable semantics to provide more machine support for user approach. The author has combined the concept of semantic web and web mining. Semantic web offers a great platform to enrich web mining. In semantic web mining the difference between content and structure mining vanishes since both are highly linked. In semantic web content and structure mining, a technique called Inductive Logic Programming (ILP) is used. The major challenges in this is the scalability of the algorithms i.e., the size of data that has to be processed and the decentralized data.[1]

2. By Anupama Prasanth / (IJETA) International Journal of Emerging Technology and Advanced Engineering.

Web Usage Mining - Its Application in E-Services.

In different fields of web mining, web usage mining has been among the rapidly growing area since it concerns about the user behavior analysis by linking the web logs with cookies and forms and exploring the access logs by unearthing user access patterns. Thus, providing user the accurate data as well as providing security to the personal data of the user. Business sector is being the most benefitted area by web usage mining by emergence of the concepts like personalized marketing, customer retention, and customer relationship. The three major tools are used to accomplish the above concepts, Preprocessing, Pattern Discovery, and Pattern Analysis. All the detailed information that are stored over distributed database can only be retrieved and accessed through web usage mining.[2]

3. By Ahmad Tasnim Siddiqui , Sultan Aljahdali / International Journal of Computer Applications.

Web Mining Techniques in E-Commerce Applications

In this paper the authors have proposed a system model that can act as a useful product in the field of e-commerce applications. The system is an integrated approach of traditional web mining technique along with e-commerce application that tends to improve the performance of e-commerce application from the visitor's point of view. The proposed system model architecture contains four approaches - Business Data, Data obtained from consumer's interaction, data warehouse and Data Analysis. The correct usage of this system model architecture comes in peer to peer applications. This paper also gives the review of the importance of semantic web and semantic web mining and its future necessity.[3]

4. By Penelope Markellou, Joanna Mousourouli, Sirmakessis Spirosis, Athanasio Tsakalidis

Using Semantic Web Mining Technologies for personalized E-learning experiences

This research paper presents the semantic web mining approach in the field of e-learning. This new emerging web technology not only has left behind the traditional methods of web mining but also have overcome many drawbacks by allowing the reuse of materials in different aspects. The technology has also provided flexible solutions, robust and scalable handling. Again to get along with the heterogeneous web resources we have come up with Ontologies. The ontology can formulate a representation of learning domain by specifying all its concepts, the possible relations between them and other properties, conditions or regulations of the domain. Hence in this paper, the author has combine the semantic web mining with the methodologies like ontologies and metadata which is further implemented in e-learning.[4]

5. By Yousef A. ALMazroui . (IJITCS) International Journal of Information Technology & Computer Science.

A Survey of Data Mining in the context of E-Learning.

The author has focused on the importance and need of knowledge management (KM). In e-learning websites the main streams of research focuses on Educational Data Mining and Learning Analytics. Another approach that author has used is the Predictive Modelling that predicts the student performance and other grace of e-learning system. There are further many issues in regard with the domain, and data privacy and ethics.[5]

6. By K. Umadevi , B. Uma Maheswari , P. Nithya / (IJIRCCE) International Journal of Innovative Research in Computer and Communication Engineering.

Design of E-Learning Application through Web Mining

The application of data mining technique over web in several useful e-services like eLearning brings the advancement in the society as well as education system. The filtration through web mining make data more personalized. The author has presented various web mining processes that can be applied to the e-learning websites to make learning more efficient both for learner as well as instructor. Associated Mining Technology, Cluster analysis, Classification. The proposed e-learning system can be divide into three platforms: Teaching resource library, which is a storage server for different data; Learning Platforms, that acts as the interactive platform for the users; and User.[6]

7. By Mohamed Koutheair khribi , Mohamed Jemni and Olfa Nasraoui

Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval

In this paper, the author has used the conceptual approach for automatic personalization methodology for appropriate filtration of data. The main objective has been given to the learners preferences and exploiting the similarities and dis-similarities among learners. The proposed framework structure of automatic recommendation e-learning platform is composed of modules: an offline module, where data pre-processing is done on the basis of learners' objectives. an online mode that provides the recommendation list by recognizing the students' needs and goals. For this the used approaches include mainly content based filtering and collaborative filtering approaches individually or in combination. The main focus is creating an e-learning platform that automatically sort the recommended approaches relying on web mining techniques and scalable search engine technology.[7]

8. By Felix Castro, Alfredo Vellido, Angela Nebot and Francisco mugica

Applying Data Mining Techniques to e-Learning Problems

The cross fertilization of data mining methods in e-learning leads to the educational data mining. The research paper surveys the data mining problems (Classification, Clustering, etc), techniques and methods (ex. Neural Networks, Genetic Algorithms, Decision Tree or Fuzzy logic). The main issues that are faced are Prediction and Visualization. In prediction techniques we need a methodology that improves the student performance and behaviour over the virtual courses in e-learning environments. For this a tool has been developed that automatically detect the typical behaviour of the students towards the web learning environment, through irregular learning processes on he basis of students response time. In visualisation technique, data exploration is to be done with regard to social network analysis adaptive to collaborative distance learning in which small learning group is measured.[8]

9. By S. Yadav, K. Ahmad and J. Shekar / IIUM Engineering Journal

Analysis of Web Mining Applications and Beneficial Areas

The author has described about different web mining processes in this research paper. The processes are divided into four first source data collection in which web resources are stored over the web server. The user behaviour is being recorded consequently over time. Next is data pre-processing that is concise to provide accuracy of data. It includes data cleaning, user identification, user session certification, access path estimation and transaction identification. The next process includes pattern discovery, where data is accessed to different pattern mining techniques, which involves path analysis, association rule discovery, sequential pattern discovery, clustering analysis and classification. The final process includes pattern analysis. Its objective is to analyse different datas through a specific pattern of model. This technique includes visualisation tools, OLAP techniques, data and knowledge queuing and usability analysis. The above techniques are highly useful in many e-services as well as e-learnin systems. [9]

10. By I-Hsien Ting

Web Mining Applications in e-commerce and e-services

Since decades, the researchers have focused mainly on the traditional web mining approach that include three sub fields of web mining technique: web usage mining, web content mining and web structure mining. The aim has been given to implement these traditional methods in such an adaptive approach that could lead to the emergence of a differently new methodology that would focus on the application over real environment. The author has given an interesting approach to the e-learning system that analysis the learning object relationship mining patterns termed as “eLORM”, that is based on learner’s access log files.[10]

3.1 Problem Formulation

E-learning system has designed as “same content”, “fits all”. But there are different people from different geographic locations enrolled as e-learners. The learning ability of students depends upon their skills as well as their background. The students’ online activities as well as course usage are significant part of e-learning systems. The study uses the techniques of data mining to understand the behavior of students using e-learning raw data. Further, Classification and clustering is done by applying different approach that involves students’ Grade as a class. The major objective of this paper is to analyze the log files of students’ raw data record of Greek university then preprocess it through data pre-processing. Various index and metric computations are used. From the study, the data mining techniques proves that there is relationship between *course usage* and *student grades*. The metric and index computation comprises enrichment and homogeneity of courses, Enrichment is a metric which express the richness of each course. It helps instructor to pay more attention to courses that have low in quality. The study uses E-learning data of the students of the Greek university, where we analyze the relation between grades and course usage of 39 courses. Various classes are used to analyze this course usage such as Enrichment, Homogeneity, Quality and Grades.

One classification algorithm is used to classify the courses with respect to grades. In this way it represents relationship between grades and course usage. Then clustering algorithm is used to cluster the classified data.

This study uses decision tree algorithm to classify the data and hierarchical clustering is used for clustering. Classification is done by taking grade as class and after classification the course usage is analyzed on the basis of grade. Courses quality depends upon course content and how often courses are updated. Enrichment and homogeneity which further results in quality of course is calculated by various mathematical formulae. Finally, final score and grade is calculated by results of computations of various course quality factors such as enrichment, homogeneity, and quality. After classification, clustering of data is done; two clusters will be made one for those courses which are highly used and one for low. In this way student’s data is clustered and results represent student’s behavior. The expected results are analyzing richness of course, how course content quality affects student’s grades.

3.2 Objectives of problem

The main objectives of the study are as follows:

- To analyze the behavior of E-learning students.
 - How dedicate the students are while enrolling in E-learning system.
 - How much time they spend in e-learning.
 - How effective they find content of courses in which they enrolled.

- To analyze the usage of online courses
 - Is there any relationship between students grades and courses usage.

- Use various data mining tools and techniques to analyze student's behavior.
 - Classify student's data to analyze their behavior. Courses are classified by taking students grade As class to find out the relationship between courses and grades. In represent how course quality Affects student's grades.
 - Cluster the classified data to categorized courses with high usage and low usage.
- To find the performance of previous approach with new tool and techniques.
 - MATLAB R2010a is used to analyze the results.

3.3 Research Methodology

The methodology consists of five steps:

1. Logging the data: The step includes the analysis of log files that are stored in the databases. These log files contains the basic information of the learner.

2. Data preprocessing: In this step, we computationally remove the noisy data such as missing values and outlier detection from the data. In this part the filtration of outlier detection of log files is done for data analysis. The generated log file is being filtered and includes the following fields:

CourseID- identification string of each field,

SessionID- identification string of each session,

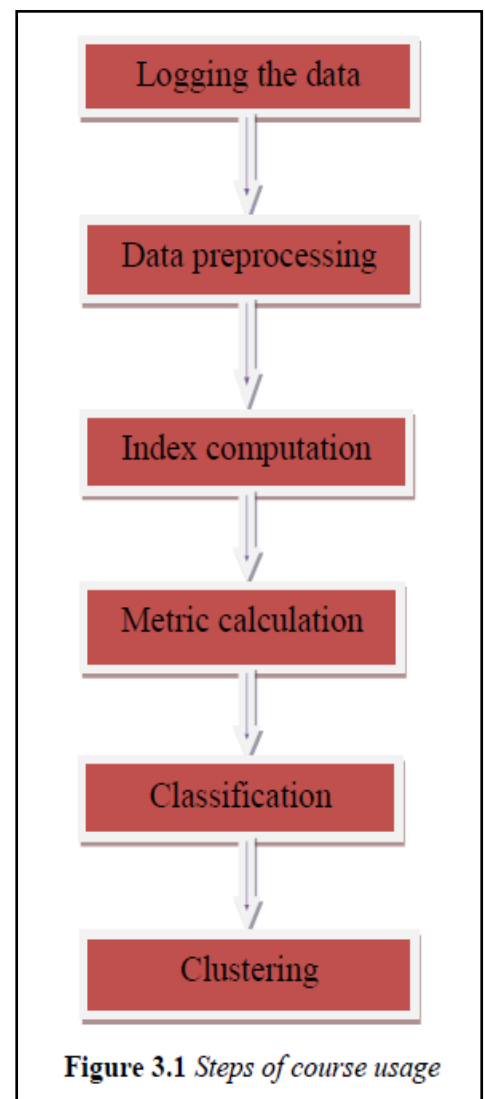
Page Uniform Resource Locator (URL) - request of each page of E-learning platform that user visited.

3. Index Computation: In this step, the Indexes such as Sessions, Pages, Unique pages, UPCS are being used.

4. Metrics Calculation: This step involves the calculations of two metrics:

Enrichment and homogeneity.

These two metrics are further used basically for the evaluation of course usage: Metric Quality, which is mean of Enrichment and Homogeneity and Metric Final, which is product of quality with UPCS. The above two metrics are used to classify the data and group the courses based on their usage. At the initial phase, we use UPCS to evaluate the courses, it is quantitative index. The courses with high value of UPCS are popular among students. In some cases where courses have same UPCS value, we use one qualitative metric: Quality, which combine the Enrichment and Homogeneity with equal weights. The final result is derived from product of Quality with UPCS.



Enrichment = $1 - (\text{Unique Pages} / \text{Total Pages})$

Where $\text{Unique Pages} \leq \text{Total Pages}$.

Enrichment values are in the range [0, 1]. When users follow unique paths in a particular course its value is 0, while in a course with minimum unique pages it is 1.

Homogeneity is another metric which represents visiting of unique pages.

Homogeneity = $\text{Unique pages} / \text{Total Sessions}$

Where $\text{Total Sessions per course} \gg \text{Unique course pages}$.

Homogeneity metric value ranges from [0, 1], when no user followed a unique path its value is 0, While 1 means that every user followed unique path.

Quality = average of enrichment and homogeneity.

Final = $\text{Quality} * \text{UPCS}$

Table: Metric name and description

Index/Metric name	Description of the index/metric
Sessions	The total number of sessions per course viewed by users
Pages	The total number of pages per course viewed by users
Unique pages	The total number of unique pages per course viewed by users
Unique Pages per Course ID per Session (UPCS)	The total number of unique pages per course per session viewed by users
Enrichment	The enrichment of courses
Homogeneity	The homogeneity of courses
Quality	The average of enrichment and homogeneity values

5. Classification: In this section, we use the classifier algorithm in-order to classify the online courses as per their metrics evaluation. To improve the contents of courses their richness must be known to us. Decision tree classification algorithm is used to classify the data.

6. Clustering: In clustering section, Hierarchical agglomerative clustering is used to cluster the courses taking grade as a parameter of class. The similar items group into one cluster and different in other. This study clusters the courses in two groups: high activity and low activity.

3.3.1 Algorithm used:

Step 1: Select the data sets file.

Step 2: Store the file.

Step 3: Detect outliers in Data i.e. Apply outlier detection process.

Step 4: Classify the data by using Decision tree classifier.

Step 5: Cluster the data using Hierarchical clustering.

3.3.1.1 Classification Algorithm:

How decision tree classification work

Tree is constructed in a top-down recursive divide-and-conquer manner

1. At start, Root contains all the training examples
2. Attributes are categorical (if continuous-valued, they are discretized in advance)
3. Examples are partitioned recursively based on selected attributes
4. Test attributes are selected on the basis of a heuristic or statistical measure (e.g. information gain)

Conditions for stopping partitioning

1. All samples for a given node belong to the same class
2. There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf
3. There are no samples left

3.3.1.2 Clustering Algorithm:

How hierarchical clustering work

Given a set of N items to be clustered, and an N*N distance or similarity matrix, the basic process of hierarchical clustering is as follows:

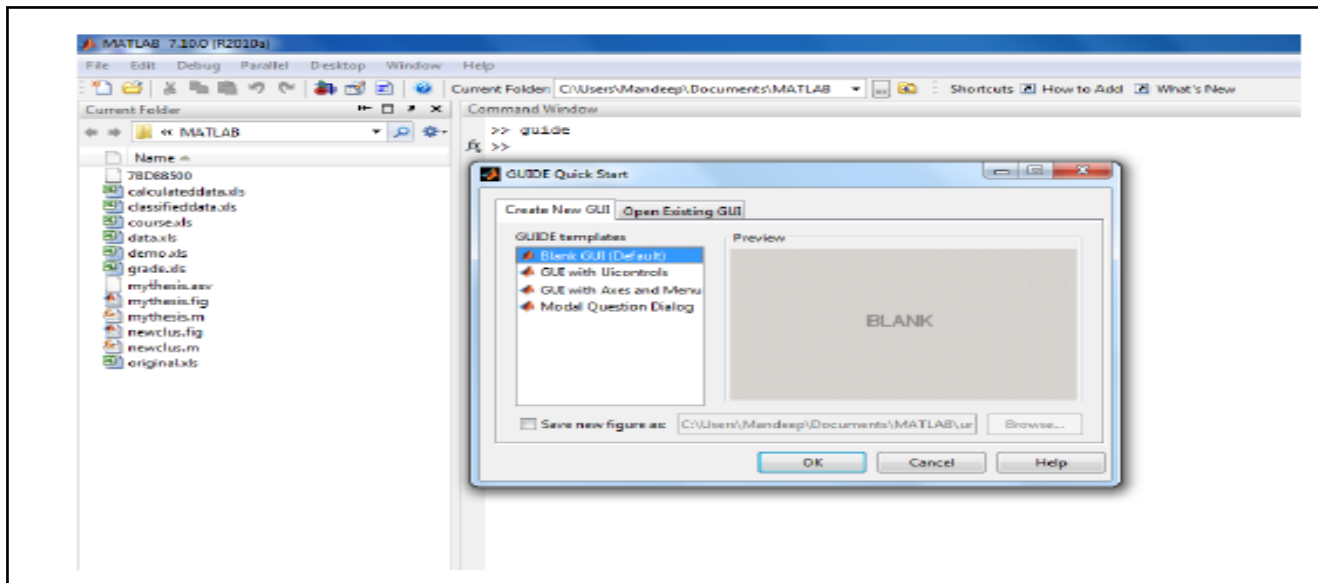
Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.

1. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
2. Compute distances (similarities) between the new cluster and each of the old clusters.
3. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

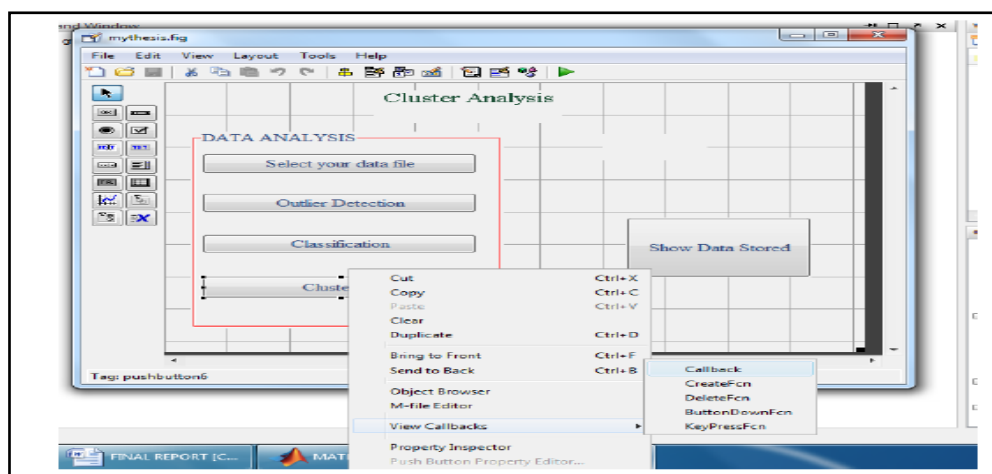
The proposed classification and clustering based methodology is implemented in MATLAB R2010a. MATLAB (Matrix Laboratory) environment is. The results are given here.

Practical Implications

Firstly we create a new project in MATLAB. Then create Graphical User Interface GUI by using ‘guide command’. Then GUI environment in form of .fig file created.



This is the .fig file which is known as graphical user interface user can add buttons, labels etc. as per its requirement. This work uses five push buttons and two labels. Push buttons for select the file, outlier detection, classification and clustering and show data storage. Labels for display the loading of data sets file. To code behind buttons we use call back function. Just right click on pushbutton and click on callback.



1. Select Data set file

Select the data file of E-learners of Greek University which is taken from previous approach. This file comprises 39 courses.

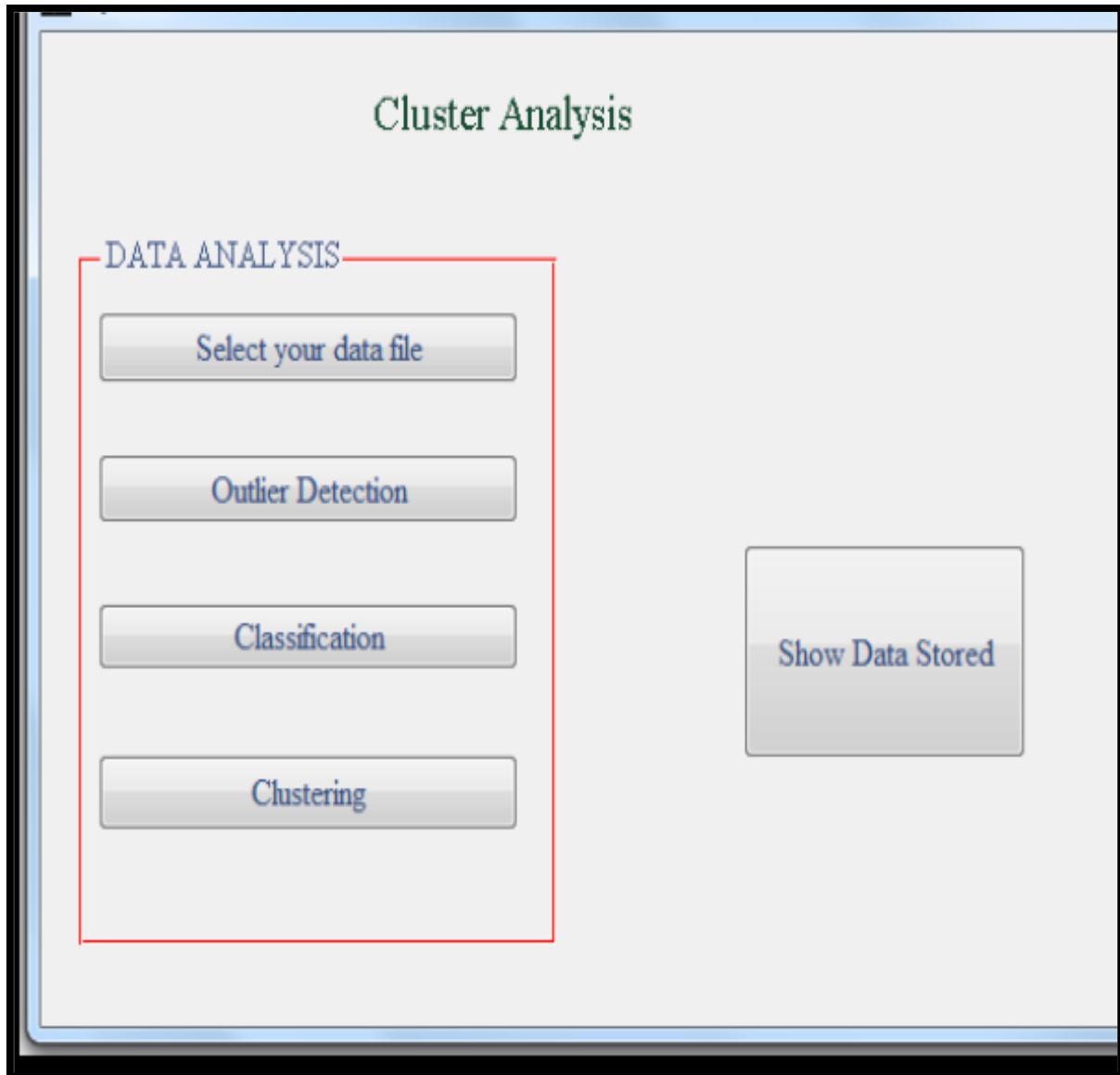


Figure *Selection of dataset file*

- Select the original data file with extension .xls.

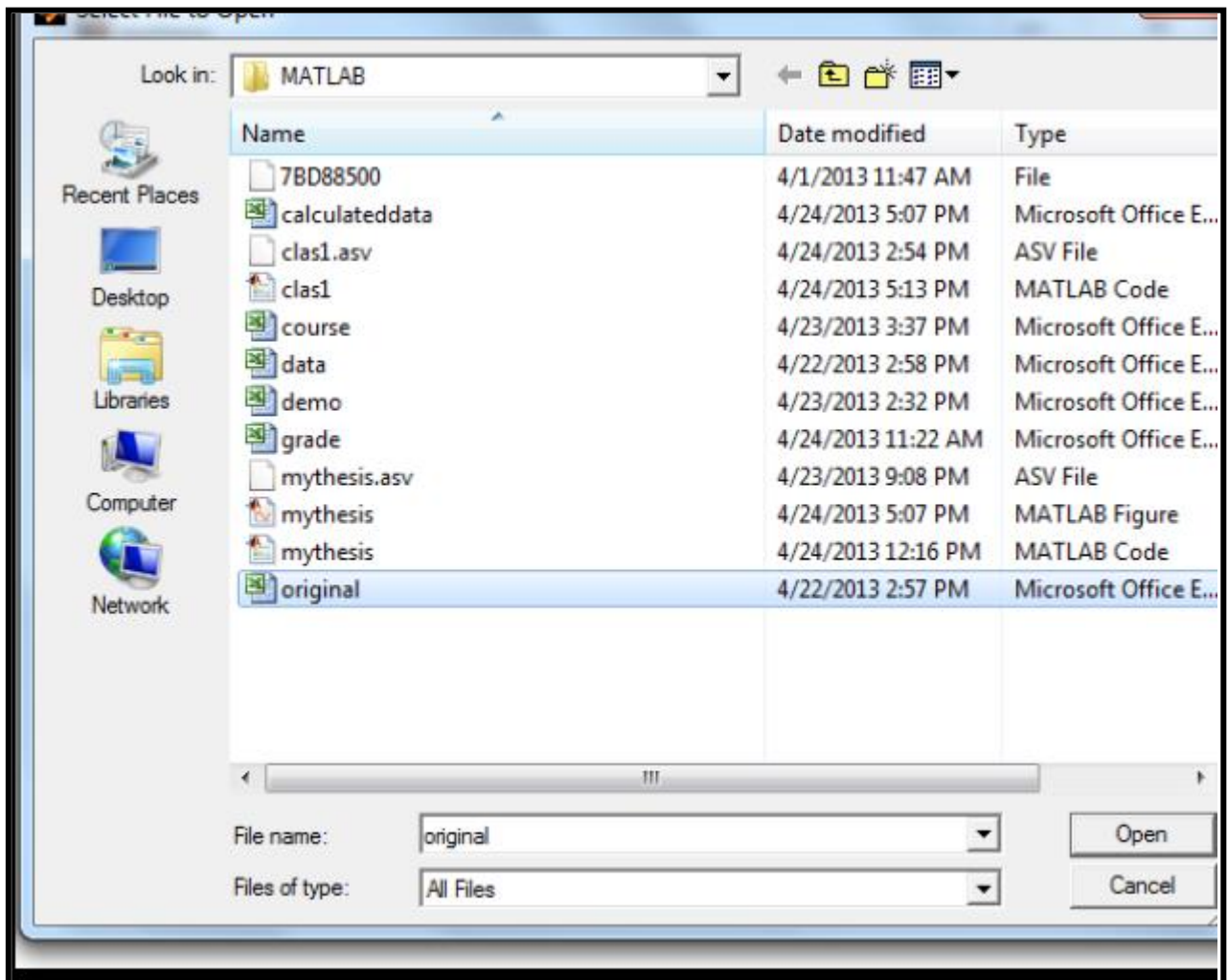


Figure. *Select original.xls*

- Wait while file is loaded.

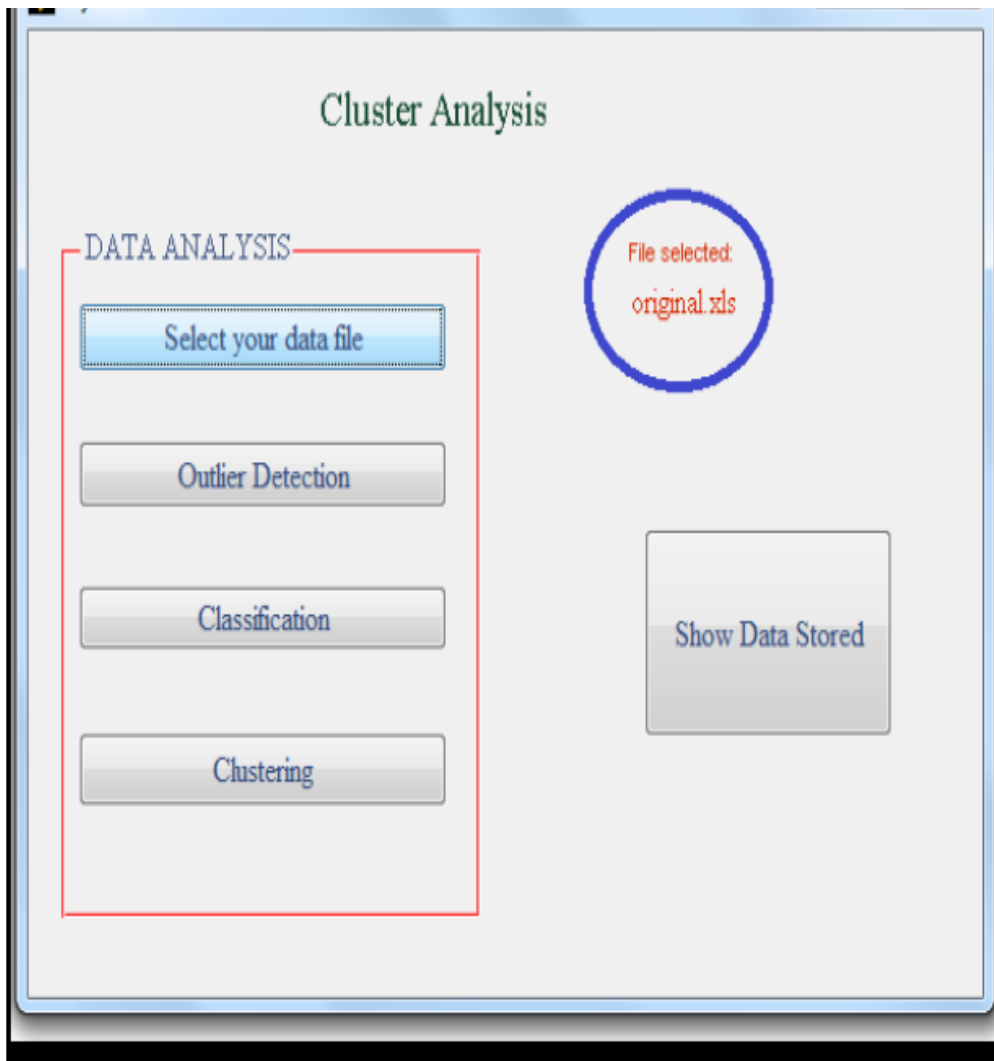


Figure. Data set file loaded

- Show stored file

Preview of C:\Users\Mandeep\Documents\MATLAB\calculateddata.xls

Worksheets

- Sheet1
- Sheet2 (Blank)
- Sheet3 (Blank)

	1	2	3	4	5
1	0.9630	0.1209	0.5419	117.0549	6.7800
2	0.9763	0.0920	0.5341	95.6115	6.1400
3	0.9696	0.0461	0.5078	93.4368	5.7500
4	0.9677	0.0972	0.5325	71.3526	6.1600
5	0.9634	0.0645	0.5140	68.8714	6.1300
6	0.9713	0.0800	0.5256	68.8596	6.2100
7	0.9568	0.0816	0.5192	66.9761	5.8700
8	0.9375	0.1607	0.5491	58.7545	5.9600
9	0.9466	0.2075	0.5771	51.3596	6.8300
10	0.9366	0.2727	0.6047	47.7692	6.9600
11	0.9407	0.1778	0.5593	45.8593	5.3400
12	0.9444	0.1176	0.5310	42.4837	6.1100
13	0.8780	0.2083	0.5432	40.7393	5.9300
14	0.9111	0.1600	0.5356	38.0244	5.9900
15	0.8784	0.2813	0.5798	37.1081	5.7800
16	0.9204	0.2813	0.6008	36.6489	6.9600
17	0.8830	0.4783	0.6806	35.3922	7.1200
18	0.9426	0.0574	0.5000	29.5000	5.8200
19	0.9125	0.2800	0.5963	32.1975	6.7500
20	0.9469	0.1579	0.5524	30.9343	6.2100
21	0.8767	0.5000	0.6884	28.9110	7.5100
22	0.8857	0.4615	0.6736	28.2923	7.1000
23	0.9032	0.2000	0.5516	25.3742	6.1800
24	0.9014	0.2121	0.5568	25.0544	6.2300

Help < Back Next > Finish Generate MATLAB code Cancel

Figure: Data sets file stored

2. Data preprocessing

This step includes removal of outliers from data. Outliers such as missing values, noise which make data noisy so cleaning the data before processing is must.

Preview of C:\Users\Mandeep\Documents\MATLAB\calculateddata.xls

Worksheets

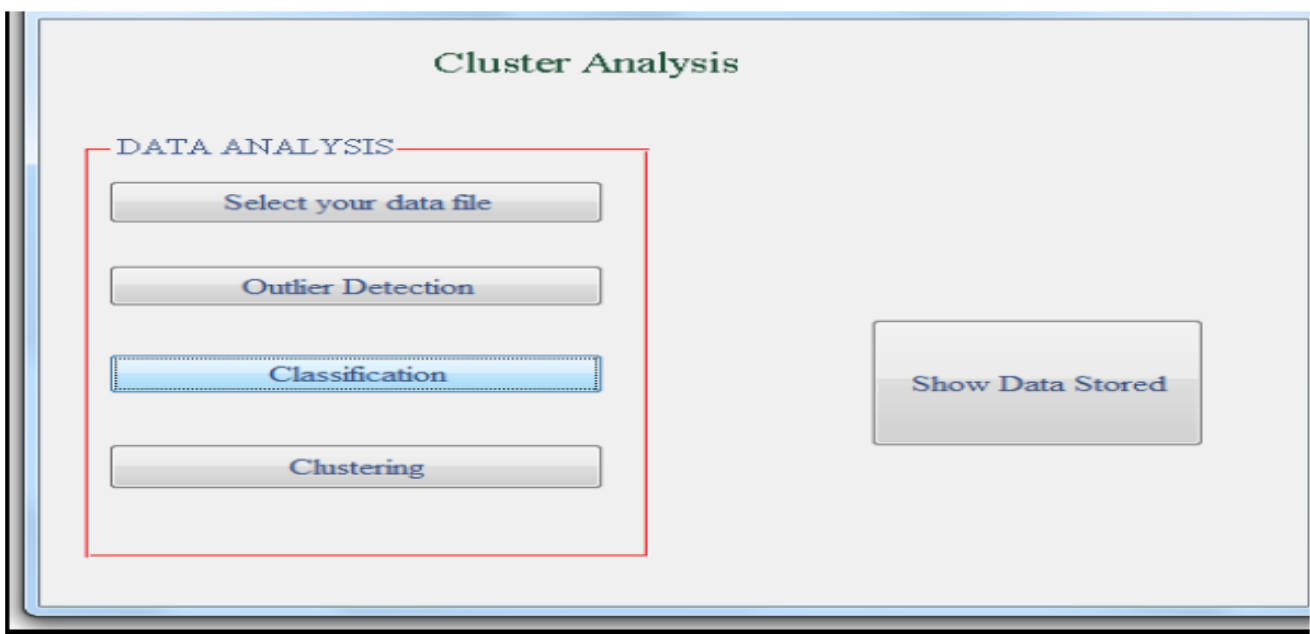
- Sheet1
- Sheet2 (Blank)
- Sheet3 (Blank)

data	1	2	3	4	5	6
15	0.8784	0.2813	0.5798	37.1081	5.7800	5.7800
16	0.9204	0.2813	0.6008	36.6489	6.9600	6.9600
17	0.8830	0.4783	0.6806	35.3922	7.1200	7.1200
18	0.9426	0.0574	0.5000	29.5000	5.8200	5.8200
19	0.9125	0.2800	0.5963	32.1975	6.7500	6.7500
20	0.9469	0.1579	0.5524	30.9343	6.2100	6.2100
21	0.8767	0.5000	0.6884	28.9110	7.5100	NaN
22	0.8857	0.4615	0.6736	28.2923	7.1000	7.1000
23	0.9032	0.2000	0.5516	25.3742	6.1800	6.1800
24	0.9014	0.2121	0.5568	25.0544	6.2300	6.2300
25	0.9625	0.0789	0.5207	23.9533	6.0700	6.0700
26	0.8511	0.3182	0.5846	22.8003	6.2300	6.2300
27	0.8158	0.5000	0.6579	22.3684	6.8600	6.8600
28	0.8837	0.2273	0.5555	22.2199	6.0200	6.0200
29	0.8478	0.3182	0.5830	22.1542	6.1100	6.1100
30	0.8689	0.4706	0.6697	18.7522	7.0300	7.0300
31	0.9048	0.1429	0.5238	17.8095	5.7900	5.7900

Help < Back Next > Finish Generate MATLAB code Cancel

Figure: Outlier detection

3. Classify the data using decision tree Algorithm.



Classification of courses is done on the basis of grades. The value of grades represents usage of courses as well quality of courses.

- Read course and grade.xls file.

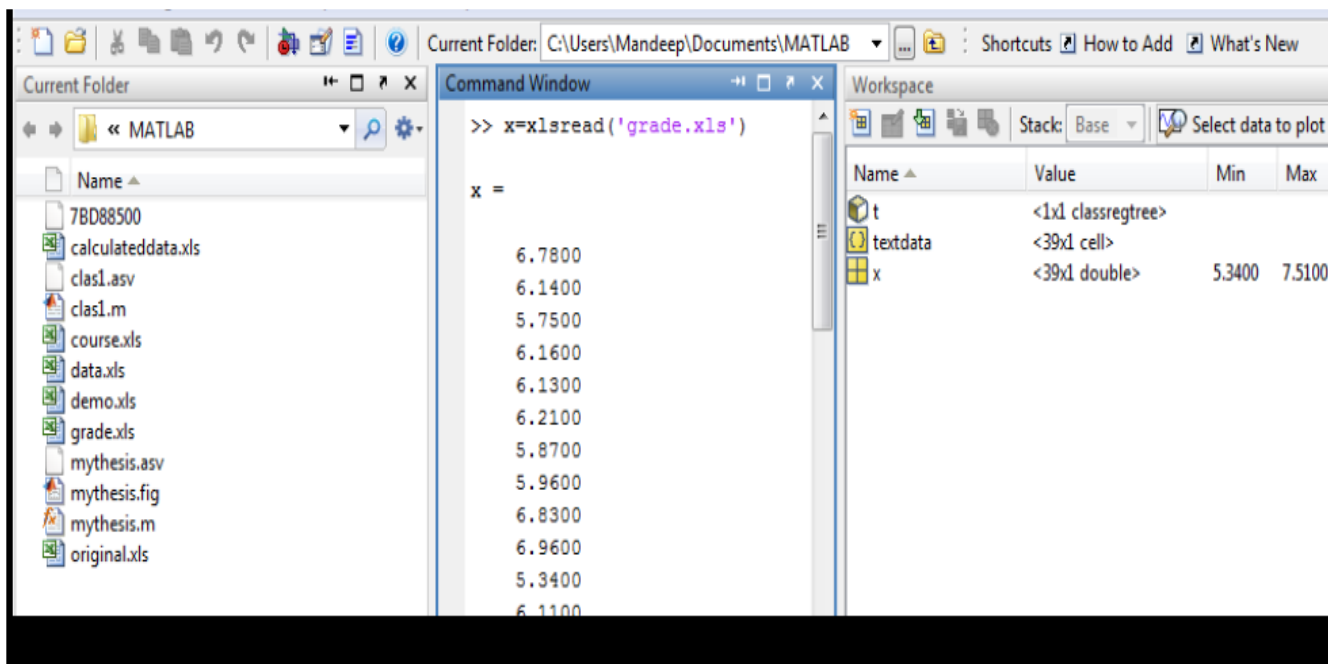


Figure. Read the grade.xls file

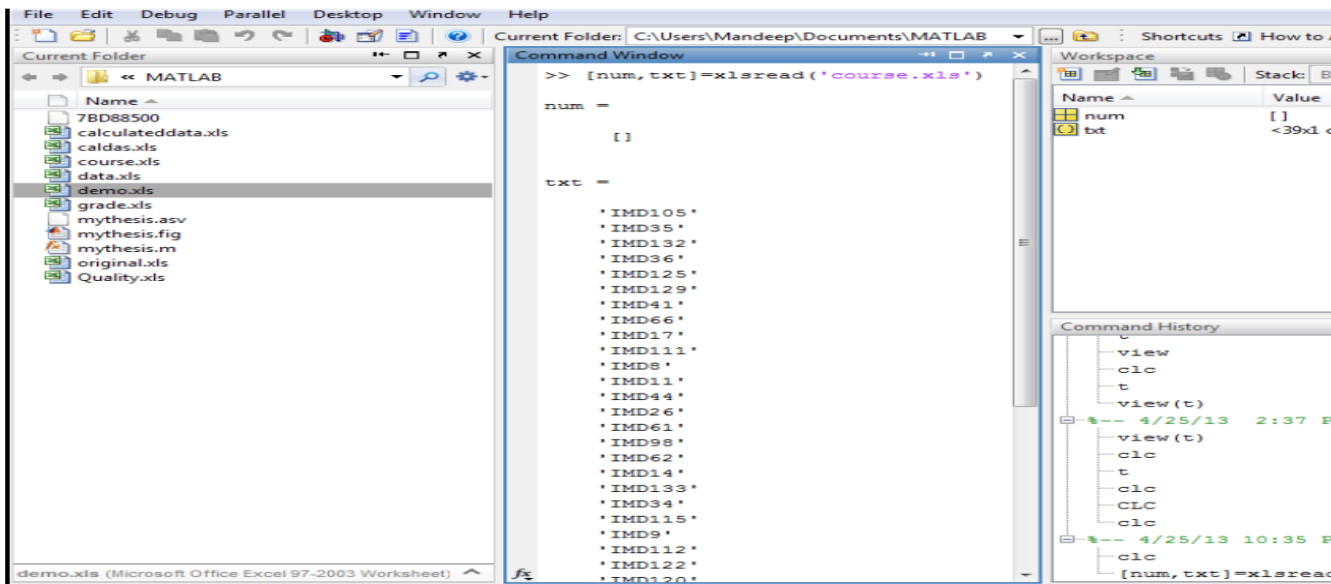
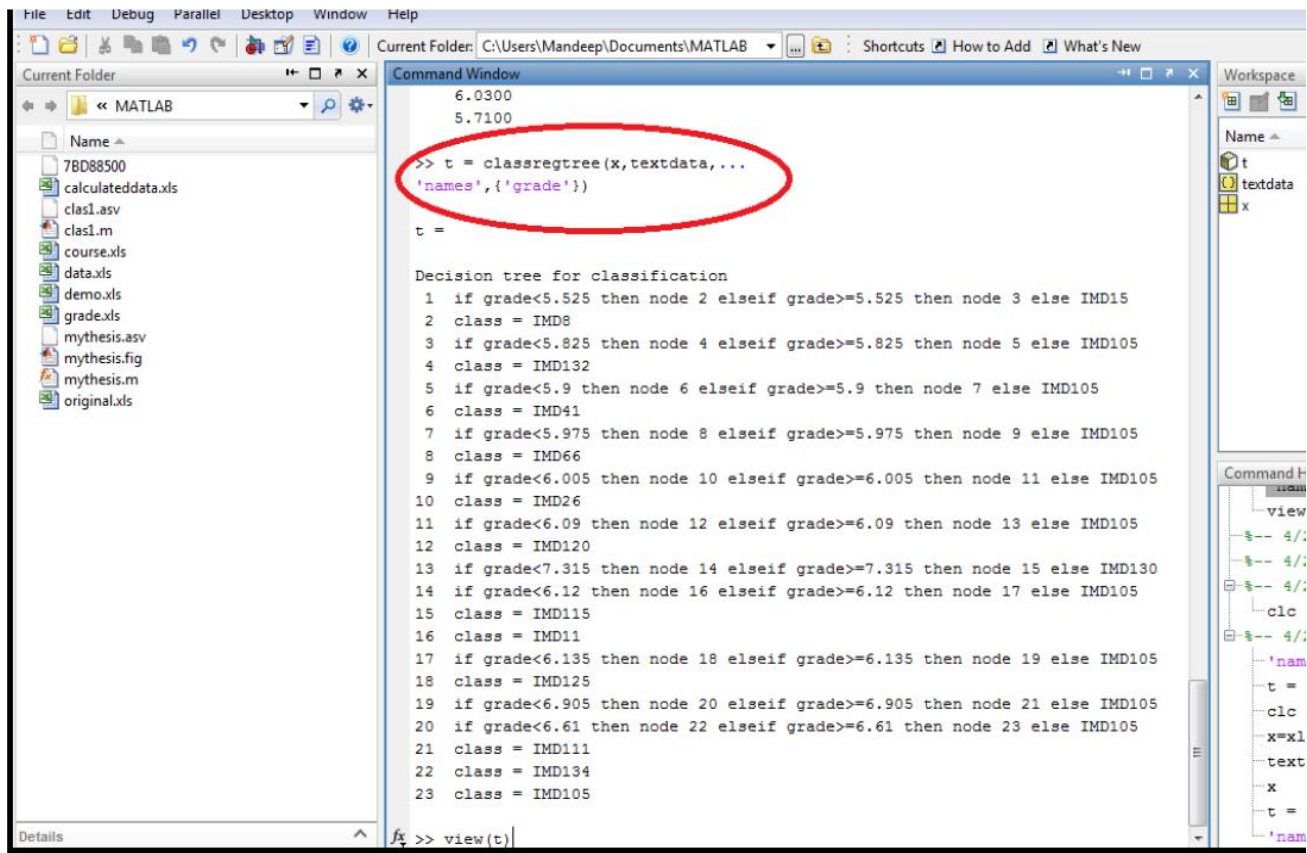


Figure. Read the course.xls file

➤ Apply decision tree Algorithm

Use MATLAB decision tree classification code and it generates decision tree classification algorithm of given data and classify the data by using algorithms' steps.



```
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\Mandeep\Documents\MATLAB
Current Folder << MATLAB
Name
7BD88500
calculateddata.xls
clas1.asv
clas1.m
course.xls
data.xls
demo.xls
grade.xls
mythesis.asv
mythesis.fig
mythesis.m
original.xls

Command Window
6.0300
5.7100
>> t = classregtree(x,textdata,...
'names',{'grade'})
t =

Decision tree for classification
1 if grade<5.525 then node 2 elseif grade>=5.525 then node 3 else IMD15
2 class = IMD8
3 if grade<5.825 then node 4 elseif grade>=5.825 then node 5 else IMD105
4 class = IMD132
5 if grade<5.9 then node 6 elseif grade>=5.9 then node 7 else IMD105
6 class = IMD41
7 if grade<5.975 then node 8 elseif grade>=5.975 then node 9 else IMD105
8 class = IMD66
9 if grade<6.005 then node 10 elseif grade>=6.005 then node 11 else IMD105
10 class = IMD26
11 if grade<6.09 then node 12 elseif grade>=6.09 then node 13 else IMD105
12 class = IMD120
13 if grade<7.315 then node 14 elseif grade>=7.315 then node 15 else IMD130
14 if grade<6.12 then node 16 elseif grade>=6.12 then node 17 else IMD105
15 class = IMD115
16 class = IMD11
17 if grade<6.135 then node 18 elseif grade>=6.135 then node 19 else IMD105
18 class = IMD125
19 if grade<6.905 then node 20 elseif grade>=6.905 then node 21 else IMD105
20 if grade<6.61 then node 22 elseif grade>=6.61 then node 23 else IMD105
21 class = IMD111
22 class = IMD134
23 class = IMD105

Workspace
Name
t
textdata
x

Command H
view
4/;
4/;
4/;
clc
4/;
'nam
t =
clc
x=x1
text
x
t =
'nam
```

Figure.Apply Decision Tree Algorithm to classify the data

Decision tree algorithm generated by MATLAB classifier:

Decision tree for classification

1 if grade<5.525 then node 2 elseif grade>=5.525 then node 3 else IMD15

2 class = IMD8

3 if grade<5.825 then node 4 elseif grade>=5.825 then node 5 else IMD105

4 class = IMD132

5 if grade<5.9 then node 6 elseif grade>=5.9 then node 7 else IMD105

6 class = IMD41

7 if grade<5.975 then node 8 elseif grade>=5.975 then node 9 else IMD105

8 class = IMD66

9 if grade < 6.005 then node 10 elseif grade >= 6.005 then node 11 else IMD105

10 class = IMD26

11 if grade < 6.09 then node 12 elseif grade >= 6.09 then node 13 else IMD105

12 class = IMD120

13 if grade < 7.315 then node 14 elseif grade >= 7.315 then node 15 else IMD130

14 if grade < 6.12 then node 16 elseif grade >= 6.12 then node 17 else IMD105

15 class = IMD115

16 class = IMD11

17 if grade < 6.135 then node 18 elseif grade >= 6.135 then node 19 else IMD105

18 class = IMD125

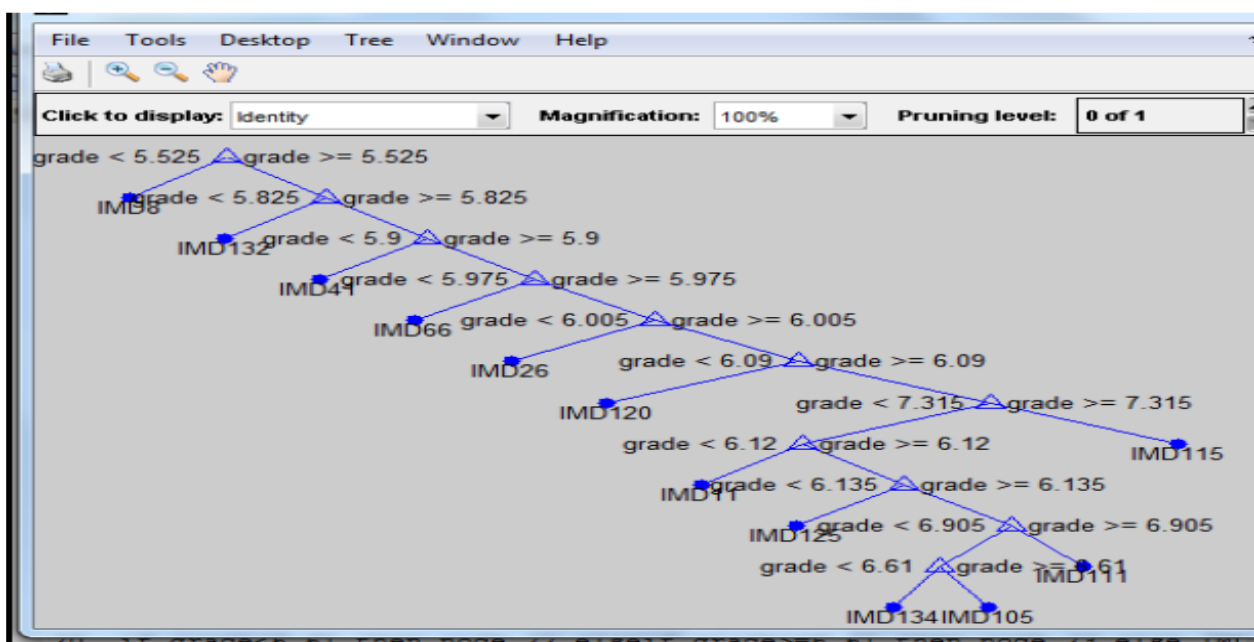
19 if grade < 6.905 then node 20 elseif grade >= 6.905 then node 21 else IMD105

20 if grade < 6.61 then node 22 elseif grade >= 6.61 then node 23 else IMD105

21 class = IMD111

22 class = IMD134

23 class = IMD105



5.1 Conclusion

The Research study proves that there is certain relation between the course usage and the students' grades. The student possess less grade because of the less richness of the course content as well as less frequent updating course contents. Therefore, due to this less activity of course usage the grades of the students are being affected. The proposed work is used with different tools and techniques of data mining to analyze the relation between course contents and students' grades. Classification of courses categorizes the rich content courses from the poor content courses, with the help of which authors can improve their course content.

5.2 Future Scope

This study calculates all the metrics and experiments of algorithms manually by using various mathematical formulae. Therefore, still some work is required so that all the computations can be done by tool. Moreover, some techniques for the automatic message of quality of course content should be there so that instructor come to know about how good or bad the course content is.

6.1 Article:

- a) <https://www.cs.purdue.edu/homes/amadkour/.../SemanticWebMining>
- b) www.scirp.org/journal/PaperDownload.aspx?paperID=26994
- c) en.wikipedia.org/wiki/Web_mining

6.2 Papers:

1. **Semantic Web Mining - A Critical Review.** By S.S. Dhenakaran, S. Yasodha / (IJCSIT) International Journal of Computer Science and Information Technologies.
2. **Web Usage Mining - Its Application in E-Services.** By Anupama Prasanth / (IJETA) International Journal of Emerging Technology and Advanced Engineering.
3. **Web Mining Techniques in E-Commerce Applications.** By Ahmad Tasnim Siddiqui Sultan Aljahdali / International Journal of Computer Applications.
4. **Using Semantic Web Mining Technologies for personalized E-learning experiences.** By Penelope Markellou, Joanna Mousourouli, Sirmakessis Spirosis, Athanasio Tsakalidis .
5. **A Survey of Data Mining in the context of E-Learning.** By Yousef A. ALMazroui . (IJITCS) International Journal of Information Technology & Computer Science.
6. **Design of E-Learning Application through Web Mining.** By K. Umadevi , B. Uma Maheswari , P. Nithya / (IJRCCE) International Journal of Innovative Research in Computer and Communication Engineering.
7. **Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval.** By Mohamed Koutheair khribi , Mohamed Jemni and Olfa Nasraoui
8. **Applying Data Mining Techniques to e-Learning Problems.** By Felix Castro, Alfredo Vellido, Angela Nebot and Francisco mugica
9. **Analysis of Web Mining Applications and Beneficial Areas.** By S. Yadav, K. Ahmad and J. Shekar / IIUM Engineering Journal
10. **Web Mining Applications in e-commerce and e-services.** By I-Hsien Ting