Student Performance Monitoring in an Online Environment

G.R.S. Perera

2017

Student Performance Monitoring in an Online Environment

A dissertation submitted for the Degree of Master of Information Technology

G.R.S. Perera
University of Colombo School of Computing

2017

UCSC

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: G.R.S. Perera

Registration Number: 2014/MIT/042

Index Number: 14550427

____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name:

____

Signature: Date:

Abstract

Nowadays, web-based educational systems are being installed more and more by universities, schools, businesses, and even individual instructors in order to add web technology to their courses and to supplement traditional face-to-face courses. Therefore identifying factors affecting the success and failure of e-learning has become essential. On the other hand systems used to demonstrate an e-learning environment accumulate a vast amount of data which is very valuable for analyzing the content of the courses and their usage from the learners has led to the deployment of data mining. Therefore the purpose of this research is to identify factors affecting the success and failure of e-learning student.

This thesis describes different data mining technologies used to identify performance factors and provide the most significant factors correlated to the success of a student in a Virtual Learning Environment. First an initial analysis of data set was done to understand the data set and based on the nature of the dataset it was decided to classify students based on their final results of Pass and Fail in the final exam which is at the end of the module. Multilayer Perception to identify factors that highly affect to the final result, CHAID Decision Tree to identify how distribution of factors affected and the relationship and direction among the factors for the final result and Binary Logistic Model to identify the probability of a student on final result and also for designing, are used as data mining techniques. Generated models are evaluated to identify the best fitting model for the dataset using classification tables, Gain charts, Lift charts, ROC curves and area under curve.

The research shows that Quizzes, Assignments and interactive media are the most influence factors while demographical factors like highest education, disability, gender and age are not to bother for a better final result. Among influenced factors number of interactions on quizzes have major impact on final result of a student. The probability of having a pass can be gain for a student can be gain in periodically and see in which factors he/she should focus on. The results and models of this study also can be used to monitor the performance of the student by giving advices in advance in order to gain better result at the end. Also designers can design their learning materials focusing mainly on enhancing the performance of the student.

Key Words: Virtual Learning Environment, data mining, monitoring

Acknowledgement

With this interim report there are so many special people mentioned below are also joined. Without their help and advice the target couldn’t be achieved at all.

While I thank all the lecturers who helped develop our knowledge, I specially thank Dr. H.A. Caldera for having faith in me to carry out the project. Your words of advice and constructive criticism make me to involve with the project in right path with motivation.

I would also like to extend my gratitude to the Open University, UK for giving opportunity for using their data set for the research purpose.

I also would like to thank the staff at Lumbini College who had the patience to bear with me when I was busy with my project.

I’m forever in debt to you all. Your unending support and encouragement helped me achieve success. Thank you all and I wish you all the very best too!

Table of Contents

List of Figures

List of Tables

List of Abbreviations

VLE	Virtual Learning Environment
LMS	Learning Management System
EM	Educational Mining
DM	Data Mining
KDD	Knowledge Discovery in Database
CMS	Course Management System
MOODLE	Modular Object Oriented Dynamic Learning Environment
CART	Classification and Regression Tree
ROC	Receiver Operating Characteristic
SQL	Standard Query Language
CMA	Computer Marked Assessment
TMA	Teacher Marked Assessment
OU	Open University
Imd	Index Multiple Deprivation
NN	Neural Network
MLP	Multilayer Perception
BLR	Binary Logistic Regression
CHAID	Chi-square Automatic Interaction Detector

Chapter 1: Introduction

Everyone has experience on traditional learning which an ancient method of learning that a teacher gathered students to a certain place a classroom, laboratory or seminars to give knowledge on a particular subject for students. This is a face to face learning. The quality of traditional learning always influences size of the classrooms (number of students) and knowledge of lecturers, its boundary to time and locations are the other weaknesses of traditional learning [1].

But today the population has increased and the education is not limited to an age boundary. Universities and other educational institutes has open their gate for more students to come up with their courses specially students who are married, having children, doing full time or part time jobs and with other responsibilities in their life.

The growth of technology with internet and its impact on education created a new model for education system called e-learning. The term e-learning interpreted differently by different researches as distance learning, online learning and networked learning. E-learning has become a common style of providing educational materials especially in universities in every part of the world as Information & Communication Technology has become a main device of the knowledge gaining. The definition of eLearning, ‘learning facilitated online through network technologies’[2], best defines online learning as it is delivered via the Internet. E-learning resources include hyper-linked and textual materials, incorporating pictures, graphics and animations, video elaborations on subject matter, just in time access to a range of electronic databases and online libraries, just in time access to e-learning communities and peers. Currently many platforms for e-learning has been established and Web2.0 is widely used. Traditionally e-learning in the higher education model, i.e. at university, has been engaged to: (1) rise visibility of university, (2) stretch the educational suggestion, and (3) make learning as visualization [3]. Furthermore e-learning is a crucial device that professors can use to enhance students' motivation and education.

1.1 Overview of Open University Virtual Learning Environment

The Open University offers a wide range of courses at various levels-from certification and diplomas to undergraduate and postgraduate degrees. Commonly students apply to study an undergraduate or postgraduate degree. For undergraduate degrees with Open University previous qualifications aren’t required while a postgraduate or master degree is undertaken after an undergraduate degree obtained. Undergraduate modules are typically six to nine months part time. Open University Qualifications are modular in structure. The credits from a module could count towards a certificate of higher education, diploma of higher education, foundation degree of honours degree.

Online programs have been designed as to be flexible in terms of learner access. It is independent for both place and time (with the exception of scheduled assignments). Learners are encouraged to be independent and to take responsibility for their own learning. Learners can access the module materials in whatever order they choose, and progress through the materials at their own pace.

The learning objectives and learning outcomes for the module are precisely stated, so that learners know exactly the degree of understanding they have to achieve and the level of knowledge required to complete the course. The module materials are presented variety of ways like text-based, text based with diagrams, interactive materials with animations, demonstrations, exercises and self-tests. Learner interacts with these materials to have the desired results.

Resources are where possible hyperlinks to electronic versions of books, journals and newspaper articles and other relevant web based resources. Learners are expected to make use of discussion forum for planned discussions and exchanged information. The tutor will provide a specific subject matter and a discussion can make on this. Learners are expected to initiate communication with the teacher or fellow learners by this and there being no requirement to do so.

The objective of the multiple choice quiz is that it allows learners to check their knowledge and understanding the course unit materials. The questions are written in clear and precise language and have a high quality content. On completion of a question learners can choose whether or not to turn to the correct answer, with its explanation of why that answer is correct. The results of the quizzes are not recorded and learners can attempt the questions as often as they like. When questions are wrongly answered explanations are provided, which allows the learner to move on to the next learning task without having to refer the course materials to correct their mistake. Quizzes are allowed for each chapter of the module and number of quizzes are relate with the tutor decision.

There are subpages that point to other sites despite of the main content of the module. For example learner will easily find a video or playlist to explore the subject matters on YouTube. There is a YouTube playlist where a learner can find videos relevant for the learning content. At the same there are some other applications such as audioBoom another way for videos, Google play for free eBooks, Open Research Online (ORO) which is one of the largest university research collection in UK that available online and FutureLearn which has a diverse selection of free, high quality online courses from some of the world's leading universities. These different sites are used for further knowledge that for who are interesting on that.

The Home page of the course allows the course materials to be flexible, and instantly recognizable to the learners. The blocks (chapters) of the module are shown in the home page and each blocks are divided into sub blocks. Learners can move to page to page within each blocks. This hierarchy of the system has a place marker so that when learners click on a sub block the name of the sub block displays on home page changes color. This gives learners a clear reference which indicates blocks they have accessed. Learner can check regularly on administrative information like registration, credits, payments and the contact information of the instructor which are visible at the home page. Also Notices in the Home page is a way of giving learners information about their course, their learning, their assessments and other university related information. Upcoming events of the module especially on assignment submission deadlines are shown in advance to the learner with quick links. Homepage can use for a learner to gain mainly with the ongoing and future updates on the module and interact with the module regularly.

There are continuous assignments and a final exam in order to assess whether the desired learning has occurred. The continuous assessments made up of tutor marked assignments (TMAs) and computer marked assignments (CMAs). CMAs are submitted via online. TMAs are usually essay or short answer questions. CMAs are made up of a series of questions and you choose the answers from a given selection. For CMAs the submission date is before the final exam date. These CMAs are work as formative assignments which set for teaching purposes only and the scores don't count to the final result of the course. These CMAs must be submitted even though the scores do not contribute to the final result. There are two different methods of submitting TMAs either online via a link or on paper. The paper TMAs can submit via post mail to the university. If TMA is not received on or before the submission date it will not be marked. The weighted average of the marks for the TMAs count for the final result. Hundred marks are allocated for each TMA. Tutor will submit the marks with comments for each TMA in home page of the module. These assignments may include only a one chapter or several chapters combined together.

Final examination is usually be handwritten and will usually three hours. The question paper is always be an unseen one, but a specimen paper will be available to familiarize with the appearance of the question paper. Also past papers are available to buy. A notification of the examination allocation will be display in module home page approximately 8 weeks before the final examination. Average marks above forty for all assessed work considered as a pass student at the end of the course.

The learner can choose to suspend or delay the completion of the study and or assessments at any point of the module but before the final examination date. If so learner can join with the next presentation to complete it. Also learner is able to carry the assignment scores that already completed to the next presentation which is referred to as Assessment Banking.

1.2 Motivation

A Learning Management System (LMS) is a software that automates the administration of learning, teaching and training events online. All LMSs manage the login of registered users, manage course catalogues, track learner activities and results and provide reports to managers. WebCT, Blackboard, Manhatten, Moodle are some examples for popular LMSs.

Most of LMSs have inbuilt tracking tools that allow them to record participation level of each student. Student tracking data captured every movement of student as they navigate through LMS. For example this data can provide lecturer or other interested parties with information when student log in, how much time they spend in the LMS, the number of messages they have read and posted, which tools and resources they have used and the number of types of files they have accessed. Tracking student online activity can provide early warning indicators of student performance [4] which is important as visual and aural cues present in face to face environment is missing in an e-learning environment. For example researches have shown that the total number of home pages visit during the first stage of study can be predictive of eventual academic outcomes in the course. A research found that discussion forum actively had a direct relationship with students' final grade [4].

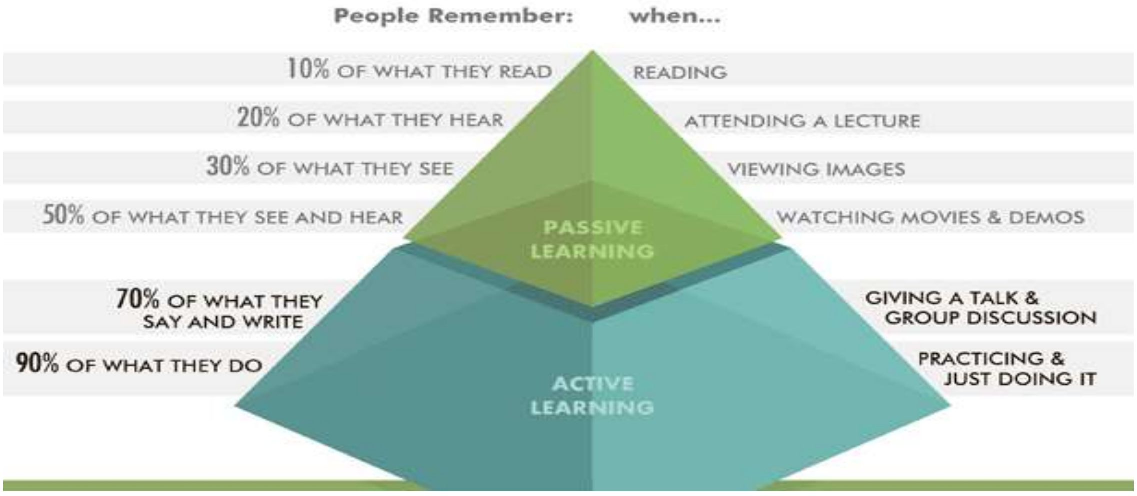


Figure 1.1: Active learning and student performance (elearninginfographics.com)

Figure 1.1 shows active learning which involves giving a talk and group description, practicing and doing it, watching movies and demos are very effective for student performance on learning. These can be gained easily from a one platform (LMSs) and also easily can accessed by any student at any time and pace.

Research have shown that some elements of student behavior at online can be predicted and that student tracking can be used to achieve both teaching and learning goals, informing ongoing evaluation, highlighting student needs, and suggesting which types of students struggle in such environment. This will ensure the performance for the student learning style as well as the design of the course. Hence these become motivational factors on working out this type of research as e-learning is becoming a most developing education system in Sri Lanka and students are now motivating to involve with e-learning systems even in schools.

1.3 Problem Definition

In modern education various information systems are used to support e-learning. Majority of these systems have logging capabilities to audit and monitor the process of learning and teaching. Learning Management Systems (LMSs) collect information about students, their enrolment in particular programs and courses, and performance like grades. These data can be analyzed and give us more insight in the overall educational system. It is possible to track how different learning resources (video lectures, handouts, wikis, hypermedia, quizzes, etc.) are used. This information readily available just in a click away. But it would equally see that unstructured information make the educational system to work as a traditional way without providing any valuable structured information on enhancement of learning and teaching process.

Student performance can be measured by a number of indicators including: successful completion of course, course withdrawals, grades, added knowledge and skill building [5]. Student performance is well understood to be a multivariable phenomenon affected by study habits, prior knowledge, communication skills, time available for study and teacher effectiveness [6].

In present environment success of a learning is defined as ensuring achievement of every learner. To reach this goal educators need tools to help them identify students who are at risk academically and adjust design strategies in virtual learning environment that help for a success learning. The instructor has to find learner's current performance level, give advices for the learner in advance on what learner should focus mainly in order to achieve a success result at the end.

However, due to the vast quantities of data these systems generate daily that is very difficult to manage data manually. Instructors and designers demand tools to assist them in this task, preferably on a continual basis. LMSs produces several kinds of reports like logs, activity report, course participation, activity completion and statistics. But it becomes hard for an Instructor to extract structured information as a summary to find the effectiveness of module and the student performance when there are huge number of students.

This research mainly focuses on allowing educators to thoroughly track and assess learners' activities while evaluating the structure and contents of the course and its effectiveness for the learning process. The goal is to identify which activities influence for the learner in order to obtain a better final result and also to identify as early as possible the students who are at risk of failing. By this the student can be helped to pass the module and the overall cost of interventions is affordable. Monitoring the performance on the success of a student is based on experience of students who are with similar characteristics in a previous run of the same module.

Use of data mining can obtain interesting patterns from large data collections. Some of the most useful data mining tasks are clustering, classification and association rule mining. In last few years research have begun to apply data mining methods to help instructors and admins to improve e-learning systems.

1.4 Aims and Objectives

1. Main objective is application of a suitable data mining methodology on measuring the performance of learning of student using the log data.

2. In order to monitor the student performance it is important to identify different factors affecting for a success academic result.

3. Acquisition of the fair knowledge on the domain and background of e-learning is also of significant importance. Detail study on the dataset obtained will be carried out.

4. Picking a suitable data mining is crucial as it will affect the interpretation of results. Sufficient research should be carried out in data mining field. Further research should be done to discover the most efficient algorithms available under the selected techniques which increase the accuracy of the prediction.

5. Build and evaluating a model to monitor the performance stage of a student

1.5 Scope

Data Mining is a well-known method for extracting pattern and associated knowledge that hid in a large data. This research uses data mining technique to monitor the students' performance while evaluating the structure and the content of the course in an e-learning environment.

The study comprises five main stages as follows:

- Problem and data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

The study starts with a problem and data understanding to find out classification or prediction. To obtain accurate data for an analysis, the problem is related to the collected data. However, the data preparation needs to be transformed for model construction and will have been iterating until the model is evaluated with satisfactory result.
Finally, the model will be deployed on the system.

1.6 Outline of the Dissertation

The report is presented into the five chapters according to the steps followed to the study to achieve project objectives. The chapters are as follows.

Chapter 2: Literature review describes the background and related work that has been done which is an important section in this project. Similar researches that has done before are discussed with this chapter. The relationship with those researches for this research, their features and steps followed, used techniques in different steps, outcomes and other important aspects are discussed in this chapter.

Chapter 3: This chapter describes the methodology and the design of the solution. The approach and steps followed in order to achieve the research objectives are described in this chapter in detail. Essential concepts and definitions used in each of steps are also described in order to understand the methodology, results and the outcomes. Using the approach and concepts described in this chapter the analysis is performed and the results and discussions are made based on them.

Chapter 4: This chapter will describe the finding of the analysis and their interpretations. The steps followed during the initial descriptive analysis, methods going to use in mining process, models going to create will be discussed in this chapter in detail. Also the model will be evaluated using model evaluating techniques and the accuracies will be discussed in this chapter.

Chapter 5: This chapter will conclude with a report with a discussion done based on the results and findings. All overall evaluation of the model will be provided while the limitations and the future works on the research will also discuss in this chapter. Finally an explanation of overall outcome of the project will be included.

Summary

Using Information Technology for teaching and learning has become a best approach for the modern learning environment and also the future learning. However student participation and performance via an online environment is still an open question. The objective of study is to investigate any relationship between the tracking data which was automatically collected in online environment by the system and performance of the student in an information technology course that was taught wholly online. The findings will lead for the preparation of students for their course materials in an effective way and for lecturers to monitor the course content and the student behavior in an online learning platform

Chapter 2: Literature Review

As explained in the 'Introduction' chapter data mining in virtual learning environment is an emerging research area. This chapter presents various research done which focus on academic performance, student behavior, study patterns and other educational related areas in VLE environment using data mining. The literature review is important to identify the research gap, understand the problem, identify methods and steps followed in similar kind of researches, different evaluation techniques used in assessing mining models, suitable technique which are to be used in this research and their pros and cons.

2.1 Data Mining, Web Mining and Web Usage Mining

Data mining (DM) is a step from Knowledge Discovery in Database (KDD) process, which is defined as a "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data" [7]. The term pattern here refers some abstract representation of a subset data of the data, that is, an expression in some language describing a data subset or a data subset or a model applicable to that subset.

Data mining efforts associated with the Web, called Web mining, can be broadly categorized into three areas of interest based on which part of the Web to mine; Web Content mining, Web Structure mining, and Web Usage Mining [8]. In Web mining, data can be collected at the server-side, client-side, proxy servers or a consolidated Web/business database [8]. The information provided by the data sources described above can be used to construct several data abstractions, namely users, page-views, click-streams and server sessions.

Web Usage Mining is defined as the process of applying data mining techniques to the discovery of usage patterns from Web logs data which to identify Web user's behavior [8]. Web Usage Mining is the type of Web mining activity that involves an automatic discovery of user access patterns from one or more Web servers.

As shown in Figure 2.1 [8] three main tasks are performed in Web Usage Mining; Preprocessing, Pattern Discovery and Pattern Analysis.

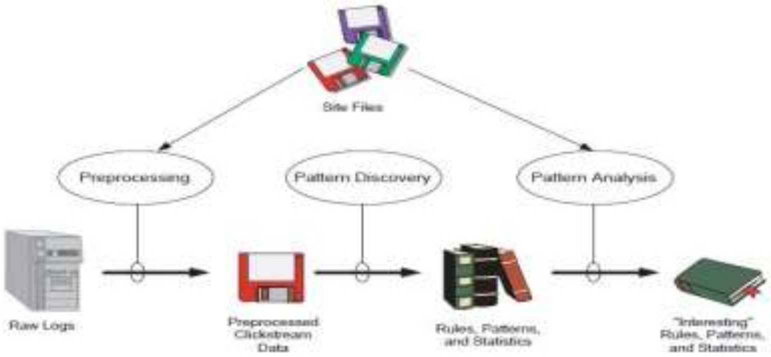
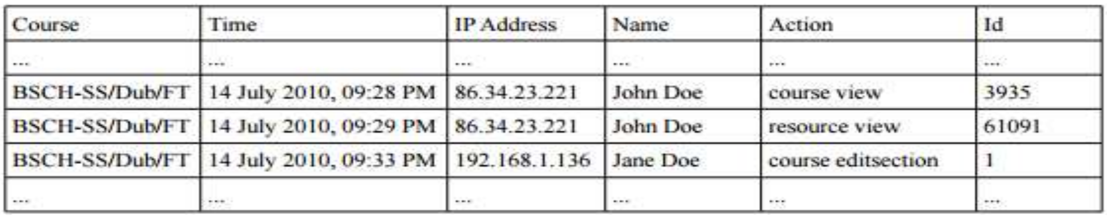


Figure 2.1: A High Level Web Usage Mining Process

2.2 Moodle

MOODLE (Modular Object-Oriented Dynamic Learning Environment) is defined as a course management system (CMS), a free, Open Source software package designed using pedagogical principles, to help educators by creating effective online learning communities [12]. This system has a number of interactive learning activity modules like forums, chats, quizzes and assignments that facilitate the learning from a participative position. In addition to these learning modules, it includes another one to register and track user’s accesses (user identification, IP and time) and the activities and resources that have been accessed. For each user, the system stores: the hour of the access, the IP address of the user’s computer, the user identification and the kind of activity or resource that has been visited [12]. Table 2.1 shows an extract table of Moodle log data.



Course	Time	IP Address	Name	Action	Id
...	...	...	...	...	...
BSCH-SS/Dub/FT	14 July 2010, 09:28 PM	86.34.23.221	John Doe	course view	3935
BSCH-SS/Dub/FT	14 July 2010, 09:29 PM	86.34.23.221	John Doe	resource view	61091
BSCH-SS/Dub/FT	14 July 2010, 09:33 PM	192.168.1.136	Jane Doe	course editsection	1
...	...	...	...	...	...

Table 2.1 Example of Moodle log data

2.3 Steps in Data mining

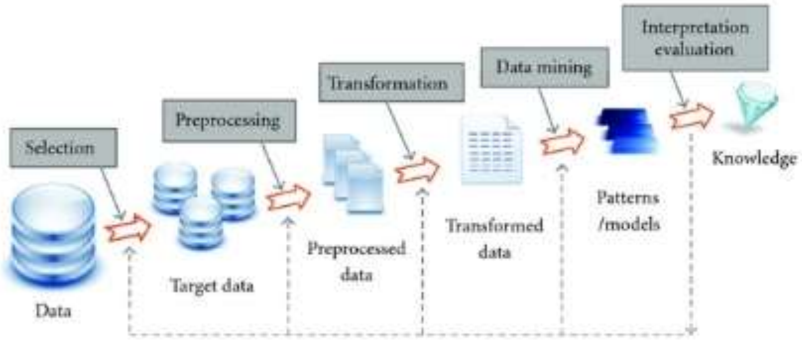


Figure 2.2 Data mining: Knowledge Discovery Process

Figure 2.2 gives a clear impression on steps of Data mining process. J. Han et al. [9] explains the process of Data mining. First the application domain should be understood well. And also should have relevant prior knowledge and the goal of the end user with the historical dataset. Then target dataset is created for discovery by selecting a dataset or focusing on a subset of variables or data samples. Pre-processing as the next step is done by cleaning the data using basic operations such as removal of noise and or outliers if appropriate, deciding on strategies for handling missing data fields, reduce the number of dimension and combine data from multiple sources into a coherent data store. Then the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve Normalization, Generalization and Aggregation. Next Data mining task can be chosen by deciding whether the goal of the knowledge discovery process is classification or regression or clustering etc. and choosing appropriate data mining algorithm. Mining on data by searching for interesting patterns of interest in a particular representational form or a set of such representation. Interpret or evaluate the mining patterns are done next and return to any of the steps above for further iteration (if appropriate).

2.4 Data collection methods

Data collection can be done in many ways where recently two methods are followed in recent researches. First one is data collection by dealing with the population directly including surveys, observations and experiments. Second method is to obtain data indirectly from data repositories where data stored for long time period. When talking about data in virtual learning environment it is usually electronic or online data collection in large data repositories where data collected and stored for a long time period [10].

2.5 Data preparation, Initial analysis and Data selection methods

Once the sample is selected, to get the better input data for the analysis a preprocessing is done in most of the data mining projects. Depending on the research objective, goal and the nature of the dataset different types of preprocessing may be required for the dataset.

C. Romero et al. [11] explained that how interaction data spread over several Moodle tables can be created as summary data files. Mainly three steps has followed on summarization the files. In the first step, which specific courses (from among all the Moodle courses) merit using mining has chosen. In the next step marks were added obtained in the final exam by the students in the course selected. And in the third step which specific attributes that need to use has to be selected.

D. Karpan et al. [12] describes how cleaning, transformation, integration and reduction were used for the dataset. Manual check of data correctness has done in order to eliminate errors. Also data format stored in database had been transformed for example: date which stored in UNIX format as it is not understandable for the researches during the correctness check. Some variables were grouped like different activities like lesson views and lesson page views on the lesson, time spent on the lesson as sum of the time spent on the lesson. Modules like blogs, chats excluded from the lesson since they were not used during the research.

In [12] it was also mentioned on different methods for treating missing values. Such values are sometimes substitute as means, constant values (ex: zero) or most frequent value. As substitute values significantly influence to the final result discarded such cases or if there are many missing values discarded the variable is preferred.

Through this literature review it is noticed that due to the incompleteness of the data set a good data cleaning and preprocessing is required in order to create an accurate data set which is to be used in the analysis phase.

Primitive data analyze is a very important step in analyzing researches which the researcher gets familiar with the data set. Having a better understandability of the data set and the nature of the attribute helps the researcher to analyze the data set with better understanding and come up with a good solution.

Usage of simple statistical tools is very helpful in primitive data analyzing phase since it helps to gain a better understanding of the data set. In many researches it can be seen that many graphical tools such as histograms, bar charts and pie charts are used to perform the initial analysis [13].

2.6 Use of Mining Algorithm

Once the preprocessing and the analyzing are done more suitable attributes can be found and these attributes can be used to build the mining model. Various types of models have been used which were selected according to the nature and the distribution of the data set and the objective of the research. Section 2.7 describes the literature review done on different models used in previous researches

2.7 Previous Research

2.7.1 Clustering

Clustering can be used to finding clusters of students with similar behavior patterns. This pattern in turn reflect a difference in learning characteristics, which may be used to give them differentiated guideline or to predict a student chance of success [14].

There are relevant cluster features that can be used to determining students' learning behaviors. A. Bovo et al. [14] used Moodle log data to show whether there is an overall ideal number of clusters that show mostly qualitative or quantitative differences.

In [14,15] Clustering algorithms provided by Weka were executed: Expectation Maximization, Hierarchical Clustering, Simple K-means and X-means. X-means has chosen the best number of clusters. But this has shown very little qualitative difference in student behavior which is a simple distinction between active and less active student. It was mentioned as explanations for this result such as small amount of dataset, the homogeneity of students and vicious circle effect.

The features that could capture of a student's online activities were mentioned by A. Bovo et al. [15] as login frequency, last login, time spent online, number of lessons read, number of lessons downloaded as a PDF to read later, number of resources attached to a lesson consulted, number of quizzes, crosswords, assignments, etc. done, average grade obtained, average last grade obtained, average best grade obtained, number of topics read, number of forums topics created and number of answers to forum topics.

Another cluster analysis were used by D. Krpan et al. [16] using Statistica to determine group of students with similar characteristics based on online and pen and paper test results, and system logs.The first step is determining the number of clusters. Statistica offers a v-fold cross-validation for that purpose.

The result was that students had lower test scores and course scores, although most of their time was spent on lessons which contain learning content. Therefore at the same time they used a statistical measurement of correlation that measures the relationship between two variables in order to determine the influence of specific variables on the test results. Both Clustering and correlations have shown same result on this. It was mentioned that such result was because some of their student do not have a computer or internet access, their learning time was organized in a computer lab and controlled by teaching assistants. Therefore it is possible that students did not well respond on a controlled environment. And also students who spent lot of time using lessons at home probably represent as outliers which could not be discarded.

M.A. Hogo et al. [17] used clustering for e-learners based on their behaviors to specific categories that represent the learner's profiles. The learners' classes named as regular, workers, casual, bad an absent. The work answer the question of how to return bad students to be regular ones. The work presented use of different fuzzy clustering techniques as fuzzy c-means and kernelized fuzzy c-means to find learners' categories and predict their profiles. Following figure 2.3 shows steps on how fuzzy logic can be applied for a web usage data set.

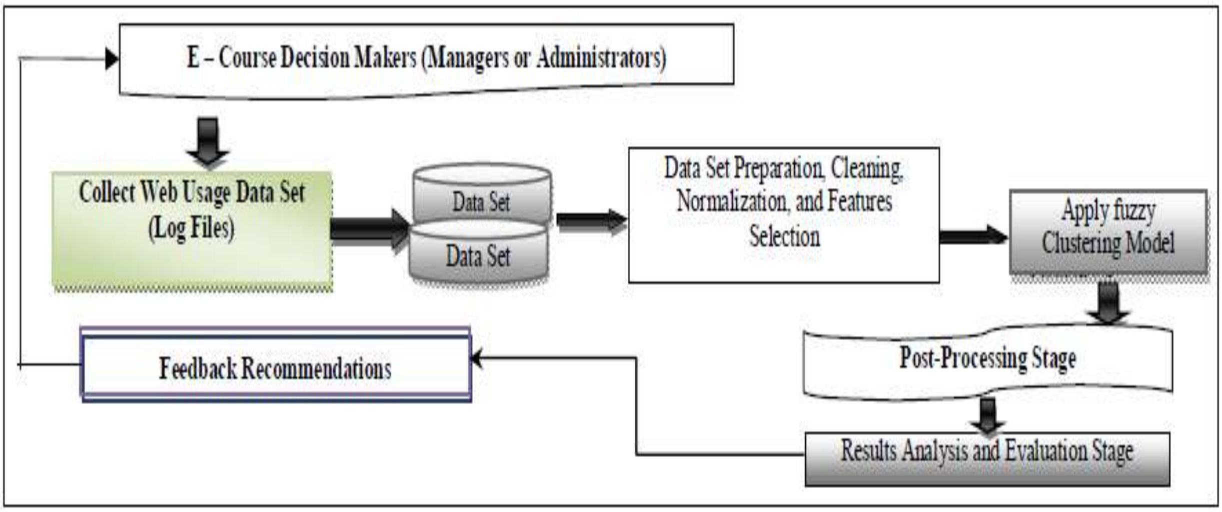

E – Course Decision Makers (Managers or Administrators)
Collect Web Usage Data Set (Log Files)
Data Set
Data Set
Data Set Preparation, Cleaning, Normalization, and Features Selection
Apply fuzzy Clustering Model
Post-Processing Stage
Results Analysis and Evaluation Stage
Feedback Recommendations

Figure 2.3 Apply fuzzy Clustering Model

A clustering algorithm was executed using a training data, after removal of the class attribute, and the mapping between classes and clusters was determined. The mapping was then used to predict class labels for unseen instances in test data. In other words, the class attribute was not used in clustering, but it was used to evaluate the obtained clusters as classifiers. This approach is shown in figure 2.4. This is used to test if student participation in forums is related to whether they pass or fail the course.

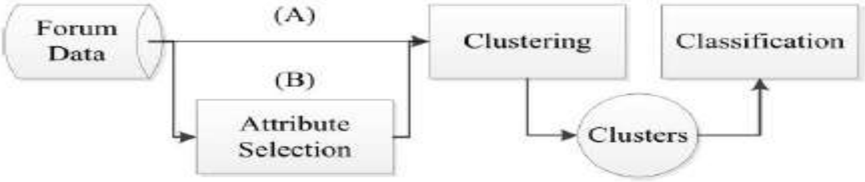


Figure 2.4 Classification via Clustering approach

2.7.2 Classification

On the other hand in educational problem it is very important for the classification model as teachers can obtain on the behavior of the student in an online environment in order to improve student learning. In general using of categorical data is much easier than numerical data as it is easy for a teacher to interpret [18].

Some experiments have been carried out by S. Ventura [18] to evaluate performance and usefulness of different classification algorithms for predicting students' usage data in e-learning system Moodle. Their objective was to classify students with equal final marks into different groups depending on the activities carried out in a web based course.

Following evaluations has been discovered [18].

- Decision tree are considered as easily understood models because a reasoning can be obtained for each conclusion. However if the tree obtained very large (more nodes and leaves) then it is not comprehensible. A decision tree can be directly convert into set of IF THEN rules which is a one of easy representation form of knowledge is. So C4.5 and CART algorithms are simple for teachers to understand and interpret.
- Rule induction algorithms also can be considered as comprehensible model because they discover a set of IF THEN rules that are high level knowledge representation and can use directly for decision making. Some algorithms such as GGP have a higher expressive power allowing the user to determine the specific formats of the rules. (number of conditions operators, etc.)
- Fuzzy rule algorithms obtain IF THEN rules that use linguistic terms that make them more comprehensible/interpretable by humans. So this type of rules can be easily understood person like teachers who are problem domain experts.
- Statistical methods and neural networks are deemed to be less suitable for data mining purposes due to lack of comprehensibility. Knowledge models obtained under these

paradigms are usually considered to be black-box mechanisms, able to attain very good accuracy rates but very difficult for people to understand. However, some of the algorithms of this type obtain models people can understand easily. For example, ADLinear, PolQuadraticLMS, Kernel and NNEP algorithms obtain functions that express the possible strong interactions among the variables.

C. Romero et al. [11] shows how web usage mining can be applied in e-learning systems in order to predict the marks that university students will obtain in the final exam. The performance of different data mining techniques for classifying students are compared, starting with the student's usage data. Several well-known classification methods have been used, such as statistical methods, decision trees, rule and fuzzy rule induction methods, and neural networks. Discretization and rebalance pre-processing techniques have also been used on the original numerical data to test again if better classifier models can be obtained.

2.7.3 Association rules

E. Garcia et al. [19] applied association rule mining to e-learning systems for traditionally association analysis (finding correlations between items in a dataset) for two points of view.

1. Help professors to obtain detailed feedback of the e-learning process:
e.g., finding out how the students learn on the web, to evaluate the students based on their navigation patterns, to classify the students into groups, to restructure the contents of the web site to personalize the courses; and

2. Help students in their interaction with the e-learning system:
e.g., adaptation of the course according to the apprentice's progress, e.g., by recommending to them personalized learning paths based on the previous experiences other similar students.

B.M. Bidgoli et al. [20] promoted in the context of web based educational systems contrast rules help to identify attribute characterizing patterns on performance disparity between various groups of students. The question was addressed using a technique called contrast rules. There was mentioned about the measurement proposed to evaluate the interestingness of association rules.

O. R. Zaane and J. Luo [21] implemented algorithm Association Rule Mining to extract useful patterns for discovering correlation between on-line learning activities with support 0.3 and confidence 0.4 which has over 200,433 entries in web log. For example an association rule looks like: 30.5% of the students who successfully finished the exercise 3 also accessed section 4 of Chapter 2.

2.7.4 Student Behavior visualization techniques

There are novel visualization techniques and tools that provides a support to teacher to observe the students' activities.

Mazza and Botturi [22] implemented GISMO a graphical interactive monitoring tool that provides useful visualization of students' activities in online courses to instructors. Moodle may benefit from GISMO for their teaching activities. GISMO provides comprehensive visualizations that gives an overview of the whole class, not only a specific student or a particular resource.c

R. Mazza et al. [23] presented MOCLog (Monitoring Online Courses with Log Files) which is a tool for the analysis and presentation of log data on a Moodle server. MOCLog is an analysis of learning activities in online-courses from a didactical point of view (learning process and outcomes), thus going beyond than simply counting and visualizing the numbers of posts and clicks.

R. Mazza and V. Dimitrova, [24] introduced a tool CourseVis a graphical monitoring tool that takes a novel approach of using Web log data generated by course management systems (CMSs). It can help instructors to quickly identify tendencies in their classes and discover individuals that might need special attention.

E. Popescu and D. Cioiu [25] introduced eMUSE (empowering MashUps for Social E-learning), which aggregates several social media components. A simple way for instructors to monitor the class activity as well as quickly check, visualize and grade each student's contributions.

Also ViSMod [26] uses concept maps to render a Bayesian student model, that exploits different types of geometric forms to represent known/unknown concepts, and KERMIT uses histograms to represent levels of a student's knowledge.

2.8 Model Evaluation techniques

When evaluating models two aspects can be considered. The first one is how to partition the data set for training and testing the models. The second one is how to evaluate and compare different models.

2.8.1 Training and testing of models

C. Mihaescu et al. [27] randomly partitioned the given data into two independent sets, a training set and a test set. Typically two thirds of the data are allocated to the training set, and the remaining one third is allocated to the test set. The training set is used to derive the classifier, whose accuracy is estimated with the test set. The estimate is pessimistic since only a portion of the initial data is used to derive the classifier.

2.8.2 Evaluation of models

Many research use different kinds of approaches to evaluate the model. One of the most basic evaluation model is the predictive accuracy. In most research predictive accuracy is calculated using a classification matrix which is also known as confusion matrix [18].

Also ROC curve are a useful visual tool for comparing two classification models [9]. Lift charts and gain charts are another method to compare mining models. They are used to measure the effectiveness of a predictive model against a random prediction and the chart provides a visual aid for measuring the model.

More details on above models are in Methodology Chapter.

Summary

In present many web based learning systems are used around the world. Therefore observation of student performance through face to face learning is a challenge now. In web based distance learning discover of access patterns to understand learners' behavior is helpful to obtain the learning objectives. In this chapter, discussions were made on some data mining steps such as data collection, preparation, descriptive analysis, data selection application of mining algorithms and evaluation techniques in previous research that could be used to enhance web-based learning environments for the educator to better evaluate the learning process, as well as for the learners to help them in their learning endeavor. The knowledge obtained in the background search was used in the next chapter, Methodology where it is further explained.

Chapter 3: Methodology

This chapter describes how this research has been conducted to accomplish the objective of the study. The chapter elaborates how the problem domain was understood, data collection methodologies, data understanding, steps followed in data preparation, how initial descriptive analysis of data was done. The steps used to design and develop the mining framework are based on the general steps followed in previous researches discussed in Literature Review Chapter. Main tools used in this research are Microsoft SQL Server 2008, Microsoft Excel 2013 and Tableau Version 10.

3.1 Understanding the problem domain

To understand the problem clearly, opinions of domain experts were used and also many literature reviews done on similar research. This study focuses on identifying which factors most influence for the student performance and among those monitoring the progress of each students. The student's performance was measured using the final result.

Most of the factors considered in this study were based on interaction of different VLE activities such as forums, home page, assignments in OuContent, Quizzes, Resources, subpages and marks for the TMA assignments of previous students. Also there are demographic factors of same students such as age band, gender, disability, previous educational qualification and number of previous attempts when starting the module

3.2 Collection of Initial Data

Due to various limitations and issues in collecting students' information in different institutes, only the Open University, UK student community was selected and hence convenience sampling was used. The Open University is the biggest university in the United Kingdom, offering several hundred distance learning courses. These can be studied both as part of a university degree and standalone modules. Students use Virtual learning systems for accessing study materials and for submitting their assignments. Initial data of the Open University, UK was collected from the web site in the format of .csv files. Two years with two semesters worth of historical data was obtained. This data source provides a unique information about student performance of their virtual learning and gives the opportunity to create new generations of learning management systems.

3.3 Description of data

Data set contains modules, students and their interactions with Virtual Learning Environment (VLE) in 2013 and 2014. Tables are connected using unique identifiers that stored in a .csv format..

3.3.1 Database Schema

Figure 3.1 shows the structure of the database schema. This database shows student demographic data with the module details and the interaction with the VLE.

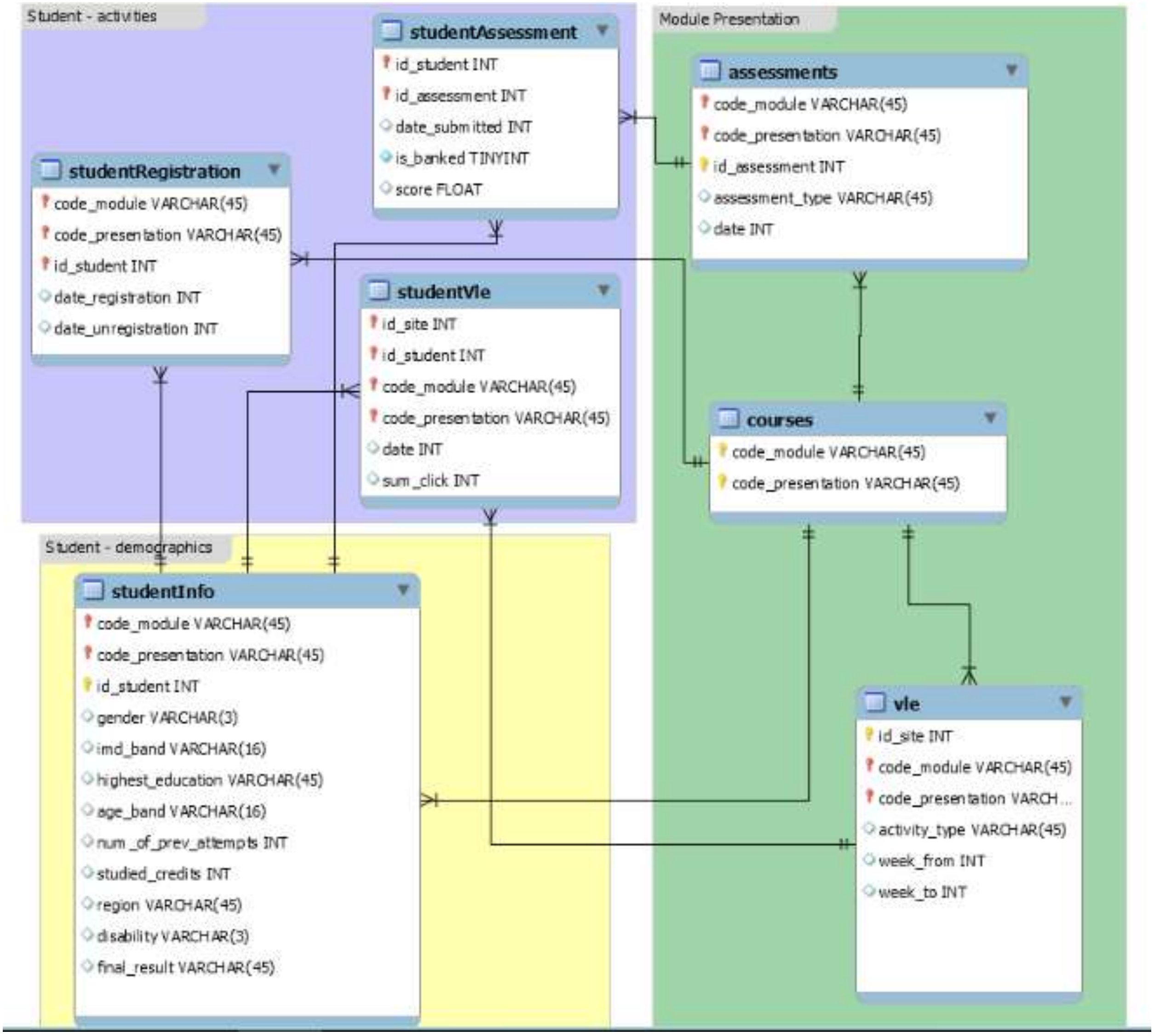


Figure 3.1 Database Schema (Knowledge media institute)

Appendix A describes the attributes available with the dataset.

3.4 Data Preparation

The output of the data preparation phase are dataset created for the training and testing purposes of the system. Few steps were followed in order to derive the desired out.

3.4.1 Selection of the potential fields

It is important to identify and select most suitable attributes to develop a module with highest accuracy.

- When allocating marks for students for each modules, assessment marks as well as the final exam have been considered.

  Two types of assessments as CMA (Computer Marked Assessments), TMA (Tutor Marked Assessments). In Open University assignments are the main way to express what have learnt. Assignments can be submitted through the online system (VLE) and some TMA assignments can be submitted on a paper. In this module no marks were allocated for CMA assignments for the final result. Anyway in order to show the performance of CMA assignments 100 marks were allocated for each. There are seven CMA assignments and the deadline of these assignments are on the final exam date. There are five TMA assignments which their marks are influenced to the final result of the students. The weighted marks of four TMA assignments are 12.5, 12.5, 25, 25 and 25 respectively. But marks for each TMA assignment were given out of 100. Therefore each assignment mark was calculated on to the weighted average mark and took total marks of four TMA assessments together which is out of 100 for each student. Among those six students have not submitted their assignments where the final result of them are fail. Assignments marks for those six students were assigned as 0. Each assignment were identified by the id number.

- The deadline date for the ongoing assessments (without final exam) were mentioned as number of days without specific submission dates. If is_bank is 1 which shows the assessment has transferred from the previous presentation shows the submission date as -1 as it cannot be indicated as 0. Therefore the submission date of assignments is unreasonable to consider as an evaluation criteria.
- The registration date of the student also shows missing values more than 50%, and from those some have missing values for unregistered date, others have minus value for unregistered date which means they withdraw from the course before the presentation of the course. Students who have plus registration days means have previous attempts

for the module. As number of previous attempts attributes is considered for the final analysis the registration dates are not much useful.

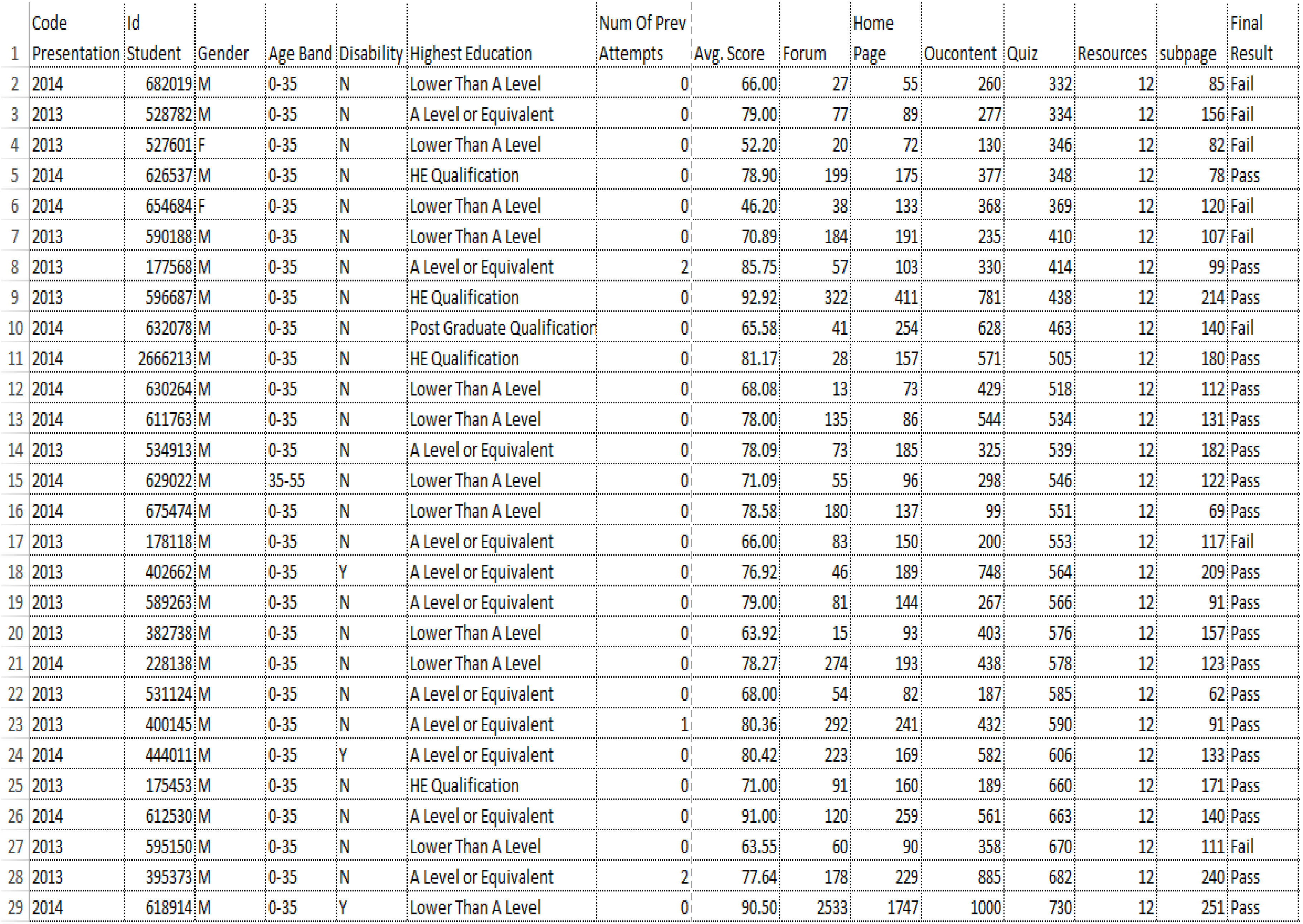
- Students who have null values for the registration date were those who unregistered before the module presentation begins, missing unregistered date and have minimum days with the module (less than 10 days). There were 45 students and their interaction with VLE is not mentioned in data set. Since this could affect to the accuracy of the dataset it was decided to remove these 45 data points. In studentInfo.csv dataset, these data been removed already other than two data points.

- Imd-band (Index of Multiple Deprivation) is a UK government qualitative study of deprived areas in English local council. Seven aspects of deprivations are covered like Income, Employment, Health Deprivation and Disability, Education skills and training, Barriers to housing and services, Crime and Living Environment. Although this index covers the majority of demographic criteria the dataset of it has significant number of missing values and the same date of 20th October 2016 has with number of data points which is not a relevant data type for this attribute.

- There were some students who failed in one presentation has faced to another presentation. Previous attempts attribute of that kind of students was calculated by gaining the previous attempts the student faced in same presentation. For example if a one student in 2013 has failed and faced again to the exam in 2014. Then his previous attempts attribute on 2014 was as 2. Therefore the student id got duplicated as it was in 2013 as well as in 2014 and it makes error on final analysis. So first attempt data were removed (data in 2013).

3.4.2 Data ready to be used

Finally a dataset with 6022 data points were considered for the analysis of the research. Figure 3.2 depicts the sample dataset with 15 number of attributes. All the attributes are explained in detail in Appendix A.



	Code	Id					Num Of Prev			Home					Final
1	Presentation	Student	Gender	Age Band	Disability	Highest Education	Attempts	Avg. Score	Forum	Page	Oucontent	Quiz	Resources	subpage	Result
2	2014	682019	M	0-35	N	Lower Than A Level	0	66.00	27	55	260	332	12	85	Fail
3	2013	528782	M	0-35	N	A Level or Equivalent	0	79.00	77	89	277	334	12	156	Fail
4	2013	527601	F	0-35	N	Lower Than A Level	0	52.20	20	72	130	346	12	82	Fail
5	2014	626537	M	0-35	N	HE Qualification	0	78.90	199	175	377	348	12	78	Pass
6	2014	654684	F	0-35	N	Lower Than A Level	0	46.20	38	133	368	369	12	120	Fail
7	2013	590188	M	0-35	N	Lower Than A Level	0	70.89	184	191	235	410	12	107	Fail
8	2013	177568	M	0-35	N	A Level or Equivalent	2	85.75	57	103	330	414	12	99	Pass
9	2013	596687	M	0-35	N	HE Qualification	0	92.92	322	411	781	438	12	214	Pass
10	2014	632078	M	0-35	N	Post Graduate Qualification	0	65.58	41	254	628	463	12	140	Fail
11	2014	2666213	M	0-35	N	HE Qualification	0	81.17	28	157	571	505	12	180	Pass
12	2014	630264	M	0-35	N	Lower Than A Level	0	68.08	13	73	429	518	12	112	Pass
13	2014	611763	M	0-35	N	Lower Than A Level	0	78.00	135	86	544	534	12	131	Pass
14	2013	534913	M	0-35	N	A Level or Equivalent	0	78.09	73	185	325	539	12	182	Pass
15	2014	629022	M	35-55	N	Lower Than A Level	0	71.09	55	96	298	546	12	122	Pass
16	2014	675474	M	0-35	N	Lower Than A Level	0	78.58	180	137	99	551	12	69	Pass
17	2013	178118	M	0-35	N	A Level or Equivalent	0	66.00	83	150	200	553	12	117	Fail
18	2013	402662	M	0-35	Y	A Level or Equivalent	0	76.92	46	189	748	564	12	209	Pass
19	2013	589263	M	0-35	N	A Level or Equivalent	0	79.00	81	144	267	566	12	91	Pass
20	2013	382738	M	0-35	N	Lower Than A Level	0	63.92	15	93	403	576	12	157	Pass
21	2014	228138	M	0-35	N	Lower Than A Level	0	78.27	274	193	438	578	12	123	Pass
22	2013	531124	M	0-35	N	A Level or Equivalent	0	68.00	54	82	187	585	12	62	Pass
23	2013	400145	M	0-35	N	A Level or Equivalent	1	80.36	292	241	432	590	12	91	Pass
24	2014	444011	M	0-35	Y	A Level or Equivalent	0	80.42	223	169	582	606	12	133	Pass
25	2013	175453	M	0-35	N	HE Qualification	0	71.00	91	160	189	660	12	171	Pass
26	2014	612530	M	0-35	N	A Level or Equivalent	0	91.00	120	259	561	663	12	140	Pass
27	2013	595150	M	0-35	N	Lower Than A Level	0	63.55	60	90	358	670	12	111	Fail
28	2013	395373	M	0-35	N	A Level or Equivalent	2	77.64	178	229	885	682	12	240	Pass
29	2014	618914	M	0-35	Y	Lower Than A Level	0	90.50	2533	1747	1000	730	12	251	Pass

Figure 3.2 Variable distribution of the selected dataset

For every student typical demographic data are collected. These include age, previous education, gender, and the number of times the student previously attempted the course. VLE data represent student’s interaction with various activity types (forum, homepage, OUcontent, quiz, resources and subpage) with the number of clicks students made on specific activity type.

3.5 Initial Descriptive Analysis

A descriptive analysis is important in order to get familiarized with the sample data. It also helps on notice missing values, outliers and the other related errors with the sample data. With the help of descriptive analysis, an understanding of the distribution could be taken. In order to display the results of descriptive analysis various kinds of visual diagrams such as bar chart, histogram, pie chart, cross tabulation, function graph, scatter plots can be used. Tableau version 10 and Excel 2013 were used for the initial descriptive analysis. Appendix C shows how different histograms helped to familiarize with the dataset.

3.6 Advanced analysis using Learning Algorithms

Upon completing the descriptive analysis the next step is to conduct the advanced analysis. Three modelling techniques that used in this study are neural network, decision tree and regression model.

3.6.1 Neural Network

Neural Network can be considered as computerized implementation of a human brain. It possesses a large number of processing units called nodes and they are interconnected by links called connections. These linked nodes process tasks in parallel to solve a problem by learning the patterns and the learned knowledge is used continuously in future problems.

In this research Multilayer Perception is used since the dependent variable Final Result is dichotomous variable (contains two values-Pass/Fail). A NN has usually three layers namely, input layer, hidden layer and output layer. The input layer contains the predictors, in this research attributes. The hidden layer contains nodes which are unobservable. Each nodes in the hidden layer is some function of the input. The output layer contains the responses, in this research final result. Each output node is some function of hidden nodes. The nodes in different layers of the NN contain connections where each connection has a weight assigned on it depending on the importance of the connection. The network learns by adjusting the weights in order to predict the correct values for the dependent variable. The function is called the activation function, and the values of the weights are determined by an estimation algorithm. The activation function links the weighted sums of nodes in one layer to the values of nodes in the succeeding layer. Hence activation functions are available for both hidden layer and output layer.

Hidden layers has two activation functions where one will be used. ‘Hyperbolic Tangent’ function transforms real values of variables to the range (-1,1). The other activation function name ‘Sigmoid’ function transforms real values of variables to the range (0,1). In this research Hyperbolic Tangent function is used.

Output layer has four activation functions namely ‘Identity’, ‘Softmax’, ‘Hyperbolic Tangent’, and ‘Sigmoid’. The identity functions does no changes to the real values and the Softmax function takes a vector of real valued argument and transforms it to a vector whose elements fall in the range (0,1). Softmax mainly use if all dependent variables are categorical. Hence in this research Softmax is used as the output layer activation method.

Additionally training can be done in three ways namely batch training, online training and mini-batch training. Online training updates the weights after every single training data record, because online training uses information from one record at a time, Mini-batch training divides the training data records into groups approximately equal in size and then updates the weights after passing the group. In batch training, updates to the weights only happens after passing all records in the training dataset. Batch training is often preferred because it directly minimizes the total error.

3.6.2 Decision Tree

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of available data.
The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset.
The CHAID which stands for Chi-square Automatic Interaction Detector is a decision tree exactly like the decision tree operator with one exception: it uses chi-squared based criterion instead of the information gain or gain ratio criteria. CHAID analysis builds a predictive model to help determine how variables best merge to explain outcome in the given dependent variable.

In this analysis categorical and continuous data can be used where continuous predictors are split into categories.

The development of the tree starts with identifying the target variable or dependent variable which would be considered as the root. Then it splits the target variable into two or more categories that are called initial or parent nodes and then the nodes are split using the statistical algorithm into child nodes. CHAID technique does not require data to be normally distributed.

CHAID's advantages are that its output is highly visual and easy to interpret. Because it uses multiway splits by default. Also it needs rather large sample size to work effectively.

Pruning is a technique in which leaf nodes that do not add to the discriminative power of the decision tree are removed. This is done to convert an over fitted tree to a more general form in order to enhance its predictive power on unseen data. CHAID uses pre-pruning. A node is only split if significance criterion is fulfilled. This reduces the problem of needing large sample sizes as the Chi-square test has only little power in small samples.

Following shows the CHAID Algorithm [27].

For each predictor variable X, merge non-significant categories. Each final category of X will result in one child node if X is used to split the node. The merging step also calculates the adjusted p-value that is to be used in the splitting step.

1. If X has 1 category only, stop and set the adjusted p-value to be 1.

2. If X has 2 categories, go to step 8.

3. Else, find the allowable pair of categories of X (an allowable pair of categories for ordinal predictor is two adjacent categories, and for nominal predictor is any two categories) that is least significantly different (i.e., most similar). The most similar pair is the pair whose test statistic gives the largest p-value with respect to the dependent variable Y.

4. For the pair having the largest p-value, check if its p-value is larger than a user-specified alpha-level α merge (alpha_merge). If it does, this pair is merged into a single compound category. Then a new set of categories of X is formed. If it does not, then go to step 7.

5. (Optional) If the newly formed compound category consists of three or more original categories, then find the best binary split within the compound category which p-value is the smallest. Perform this binary split if its p-value is not larger than an alpha-level α split-merge (alpha_spli-merge).

6. Go to step 2.

7. (Optional) Any category having too few observations (as compared with a user-specified minimum segment size) is merged with the most similar other category as measured by the largest of the p-values.

8. The adjusted p-value is computed for the merged categories by applying Bonferroni adjustments.

Splitting

The best split for each predictor is found in the merging step. The splitting step selects which predictor to be used to best split the node. Selection is accomplished by comparing the adjusted p-value associated with each predictor. The adjusted p-value is obtained in the merging step. 1. Select the predictor that has the smallest adjusted p-value (i.e., most significant).

2. If this adjusted p-value is less than or equal to a user-specified alpha-level α split (alpha_split), split the node using this predictor. Else, do not split and the node is considered as a terminal node.

Stopping

The stopping step checks if the tree growing process should be stopped according to the following stopping rules.

1. If a node becomes pure; that is, all cases in a node have identical values of the dependent variable, the node will not be split.

2. If all cases in a node have identical values for each predictor, the node will not be split.

3. If the current tree depth reaches the user specified maximum tree depth limit value, the tree growing process will stop.

4. If the size of a node is less than the user-specified minimum node size value, the node will not be split.

5. If the split of a node results in a child node whose node size is less than the user specified minimum child node size value, child nodes that have too few cases (as compared with this minimum) will merge with the most similar child node as measured by the largest of the p-values. However, if the resulting number of child nodes is 1, the node will not be split.

3.6.3 Regression

Many types of regression techniques are available to estimate the relationships among variables depending on the nature and number of the independent and dependent variables.

Among those Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous (binary) variable in which there are only two possible outcomes. Logistic regression is simply a nonlinear transformation of the linear regression.

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous (dependent) variable and a set of independent (predictor) variables. Logistic regression generates the coefficient of a formula to predict the probability of presence of the dependent variable. It contains the estimated probabilities to lie between 0 and 1.

The logistic regression coefficients are the coefficients b_0, b_1, b_2,…b_k of the regression equation.

$$\ln[(P/(1-P)] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

$$[(P/(1-P)] = \exp (b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k)$$

where,
ln is the natural logarithm, $\log_{exp}$ where exp=2.71828…
P is the probability where the event of dependent variable occurs.
P/(1-P) is the odd ratio
ln[(P/(1-P)] is the log odd ratio, or logit

An independent variable with a regression coefficient not significant ($p>0.05$) can be removed from the regression model.

For instance, the estimated probability is,

$$p = \frac{1}{1+e^{-logit(P)}}$$

Alternatively, logit table also can be used.

3.7 Evaluation of the model

It is necessary to evaluate the accuracy of the selected model. Different methods are used for model evaluation in the study which are explained in this section.

3.7.1 Training and testing of model

When applying the selected model these datasets were divided as the training and the testing set. The 70% of selected dataset is used as training dataset and else as testing dataset.

3.7.2 Classification rules

Classification tables are used to provide a summary of predicted values and the actual values. It also can be used to cross check the predicted values against actual values and prediction accuracy can be calculated using them. A classification table can be implemented as shown in table 3.2 [9].

Predicted class

		True	False
Actual Class	True	TP	FN
	False	FP	TN

Table 3.2: A sample classification table

TP - True Positive, correctly diagnosing a True case as True

TN - True Negative, correctly diagnosing a False case as False

FP - False Positive, Incorrectly diagnosing a case as true when its True state is False

FN - False Negative, incorrectly diagnosing a case as False when its True state is True.

Based on the classification matrix accuracy of the model can be calculated as follows:

$$\text{Accuracy} = (TP+TN) / N$$

when N is the total number of cases used for prediction. (N = TP + TN + FP + FN)

3.7.3 Receiver Operating Characteristic curve

ROC curves are mainly used to examine the performance of the models. An ROC curve shows the trade-off between the true positive rate or sensitivity (proportion of positive tuples that are correctly identified) and the false-positive rate (proportion of negative tuples that are incorrectly identified as positive) for a given model [9]. It displays the Specificity on the X axis and the Sensitivity on the Y axis.

Another important value used with the ROC curve is the Area Under Curve (AUC). AUC is the measure of the ability of the diagnostic tests to correctly identify the cases. Diagnostic tests with the higher AUCs are generally better and should always be higher than 0.5. The closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy will have an area of 1.0 [9].

Summary

This chapter describes the approaches taken, steps followed and concepts used in the study in order to achieve the goal. The knowledge gained by the background research work discussed in Chapter 2 was used in this chapter to design the methodology and most appropriate techniques for this study has been selected and explained. The methodology described in this chapter is followed and by using the concepts discussed here the results are obtained which will be discussed in the next chapter.

Chapter 4: Data analysis, Results and Discussions

This chapter presents and interprets the findings of the initial descriptive analysis and the advanced analysis done by applying the mining algorithms such as Decision Tree which was conducted by following steps explained in the previous chapter Methodology. It also elaborates the experiments conducted and analysis on results. Discussion on results set obtained is presented and finally an overall discussion on the entire result is presented. Accuracy levels of different algorithms used for advanced analysis is also discussed in this chapter and finally an overall model evaluation is discussed.

4.1 Initial Descriptive statistics

Altogether there were 6022 respondents in the sample considered in this study. Thirteen different variables were there in the data set. A categorical data analysis was conducted to familiarize with the data set. Since objective of this research is to identify the patterns that affect the final result which used to measure the performance of the student, the analysis was conducted with respect to the final result.

Initial descriptive analysis were presented in the following order to explain the distribution of dataset according to different attributes. Some statistics were used to explain the relationship among two or more variables.

4.1.1 VLE Activity types wise distribution

Forum, Home page, OuContent, Quiz, Resources, Subpage were used as VLE activity types on the sample dataset. Average click of each activity types of the entire sample dataset was calculated as the following frequency table 4.1.

Table 4.1 Frequency Table for the use of e-learning features wise distribution

VLE activity types	Rate
Forum	13.94%
Home Page	14.2%
OuContent	38.3%
Quiz	22.7%
Resources	1.4%
Subpage	9.5%

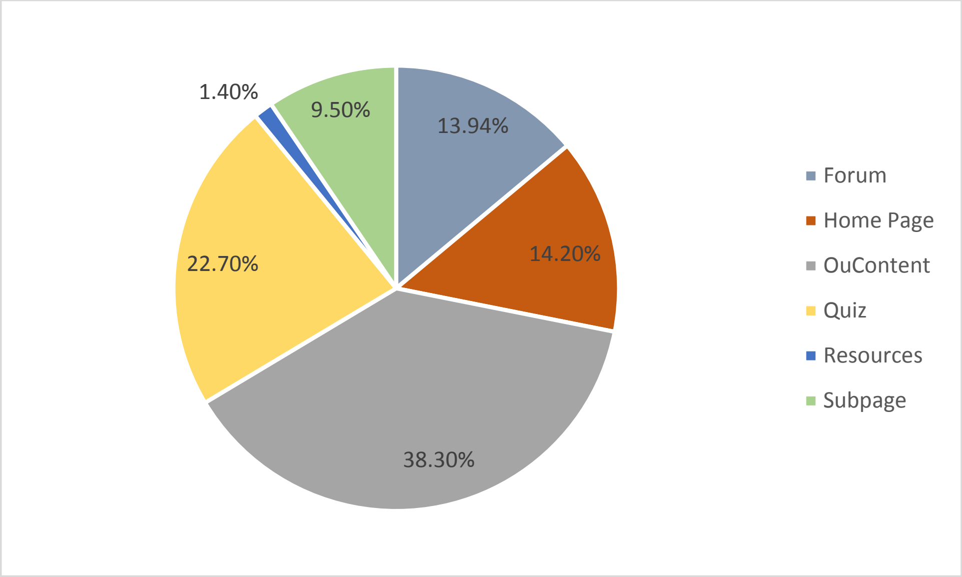


Figure 4.1 Pie chart for use of VLE activity type wise distribution

Figure 4.1 pointed out that approximately a student has 38% used of OuContent which represent content of assignments. 27.7% of e-learning resources used by a student as quizzes of this course. Home Page and Forums have nearly similar percentage of use which are 14.2% and 13.94% respectively. Subpage which points to other sites in the course were used less when compared to the other e-learning resources available. But the usage percentage of Resources of the course which usually contains hyperlinks to book, journals and articles is so low which has only 1.4% of used.

4.1.2. Final result wise distribution

Here the analysis was done to identify the pass rate. This refers to whether the student have completed the course successfully or not. The frequency table and the bar chart shown in table 4.2 and figure 4.2 respectively illustrates the percentage of the final result of entire sample data.

Table 4.2 Frequency table with status wise data distribution

Final Result	Rate
Pass	60.6%
Fail	39.4%
Total	100%

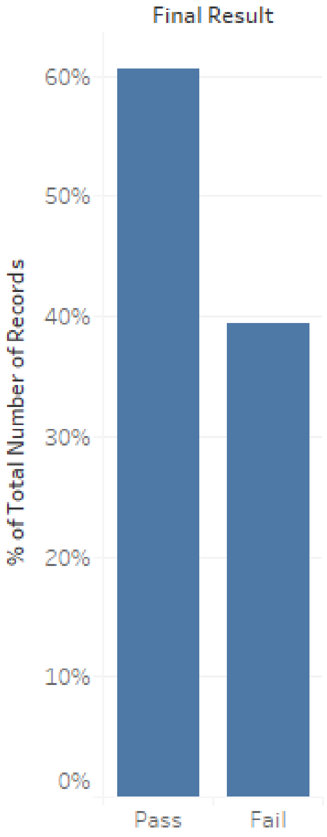


Figure 4.2 Bar chart for Final Result wise data distribution

According to the bar chart it can be see that more than 60% of students have passed while 40% of student were not able to complete the course successfully which is a considerable amount. So it is an essential to consider the reasons on the 40% of fail rate.

Usually a VLE mainly dependable on different activity types (mentioned previously) used on it to deliver a module. Therefore interactions on these activity types need to be considered. Table 4.3 is used to find if this interactions affected to complete the model in successful way. For that the number of clicks were divided in to different ranges to find whether the increase of number of clicks on activity types affected to the final result.

Table 4.3 Frequency Table for Final Result distribution with VLE activity types

Average no. of clicks on VLE activity types	Pass Rate
Click < = 50	1.85%
Click 51 – 100	16.79%
Click 101 – 800	64.21%
Click > 801	72.63%

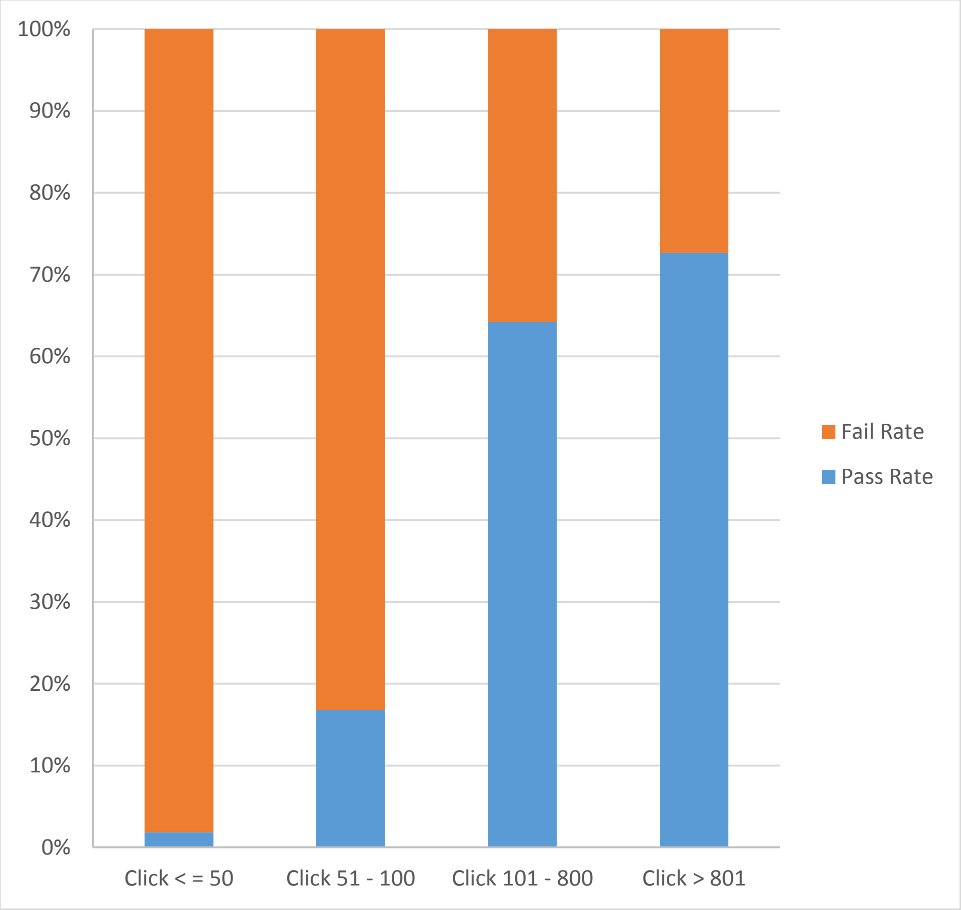


Figure 4.3 Component Bar Chart for Final result distribution with VLE activity types

The above chart shows that percentage of Pass rate gets high with the increase of number of clicks on VLE activity types. When the average clicks of VLE activity types are less than 50 the pass rate is too low (1.85%) but on the other hand high active students on VLE activities have greater probability to get a pass as the final result. When the number of clicks are in the range of 101-800 the pass rate gets increase in highly.

4.1.3. Previous Educational Level Distribution

It is interesting to get an idea about the pass rate with respect of the previous highest educational level. The idea behind analyzing the previous highest educational level and the pass rate is to check whether the pass rate depends on the previous highest educational level when enroll to the module.

Therefore first consideration was given to identify the variation between the previous educational levels of student.

Table 4.4 Frequency Table with previous highest educational level wise data distribution

Previous Highest Educational Level	Rate
A/L or equivalent	47.03%
Lower than A/L	40.58%
Higher Educational Qualifications	11.10%
No formal Education	0.93%
Post Graduate Qualification	0.36%
Total	100%

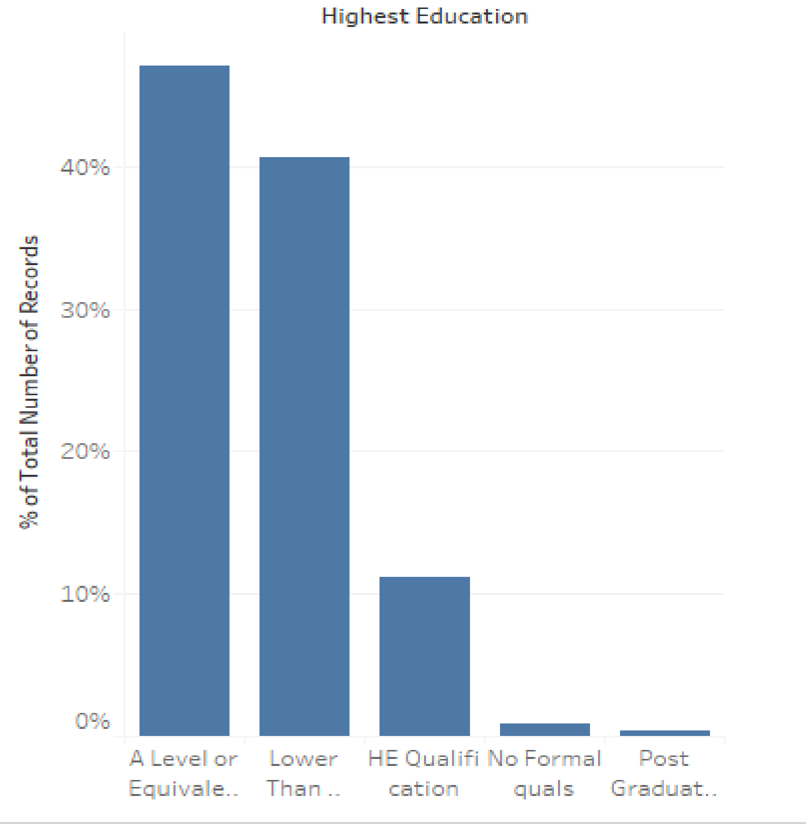


Figure 4.4 Bar chart for previous highest educational level wise data distribution

According to the distribution of highest educational level shown in figure 4.4 , students with Advanced level (A level) or equivalent has the highest percentage (47.03%) compared to the other educational levels. There are significant amount of the students whose educational level is lower than A level (40.58%). Compared to these educational levels there are few students who are with higher educational level (11.10%). Students with no formal educational level and Post graduate level are very few compared to the other educational levels.

Following frequency table shows the how final result changes with each educational levels.

Next consideration was given to identify how these different levels of educational backgrounds affect for the final result of the module. Table 4.5 was used to analyze the pass rate with the previous education qualification gained when register to this model.

Table 4.5 Frequency Table with the final result against educational level wise data distribution

Educational Qualification	Pass Rate
A level or equivalent	65.88%
Lower than A Level	49.055%
Higher Educational Qualification	68.7%
No formal Education	51.23%
Post Graduate Qualification	62.5%

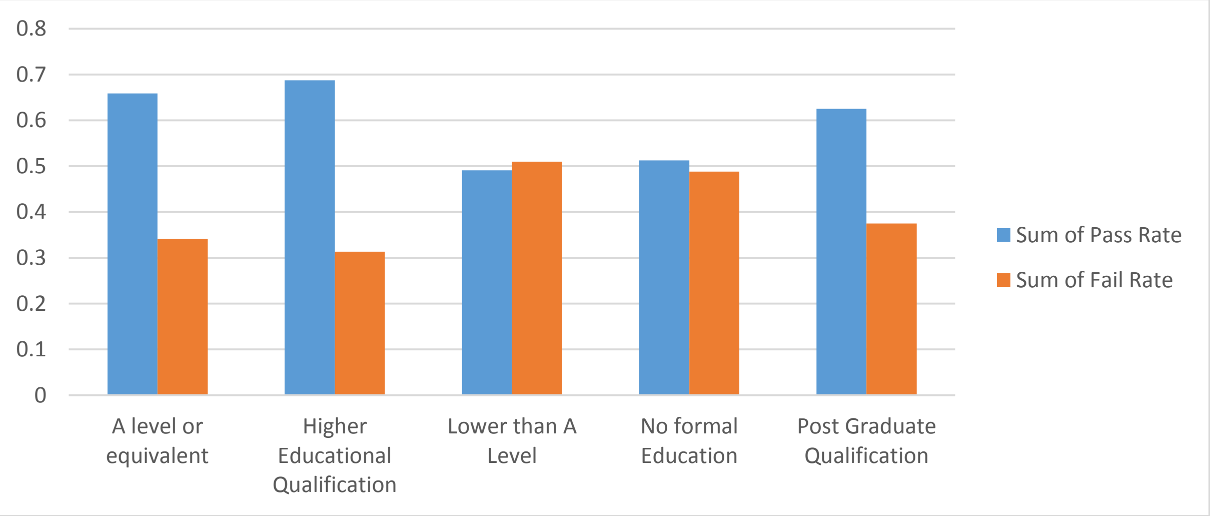


Figure 4.5 Clustered bar chart for the pass rate of against educational level wise data distribution

By figuring out the above chart it shows that students with A level or equivalent, Higher Educational Qualification and Post Graduate Qualification have high pass rate nearly 60%. But the students with Lower than A level and Without Formal Education the Pass rate and the Fail rate is not much vary which are 49% and 51% respectively. It seems nearly 50% of students without formal education and Lower the A level get fail. Highest fail percentage is with the student whose educational level is lower than A level.

In the above distribution VLE activity types not taken into account. Following frequency table illustrates how VLE activity types with educational qualification influence to the final result.

Table 4.6 Cross Table for ranges of clicks on VLE activities with Educational Level against Final Result

	Ranges of clicks on activity types	Pass Rate
A/L or equivalent	Click < = 50	1.90%
	Click 51 - 100	15%
	Click 101 - 800	64.73%
	Click > 801	71.69%
HE qualification	Click < = 50	4.55%
	Click 51 - 100	21.88%
	Click 101 - 800	66.90%
	Click > 801	70.78%
Lower Than A/L	Click < = 50	1.16%
	Click 51 - 100	16.43%
	Click 101 - 800	62.80%
	Click > 801	74.30%
No formal Education	Click < = 50	-
	Click 51 - 100	51.52%
	Click 101 - 800	60%
	Click > 801	80%
Post Graduate Qualification	Click < = 50	-
	Click 51 - 100	-
	Click 101 - 800	69.44%
	Click > 801	76.47%

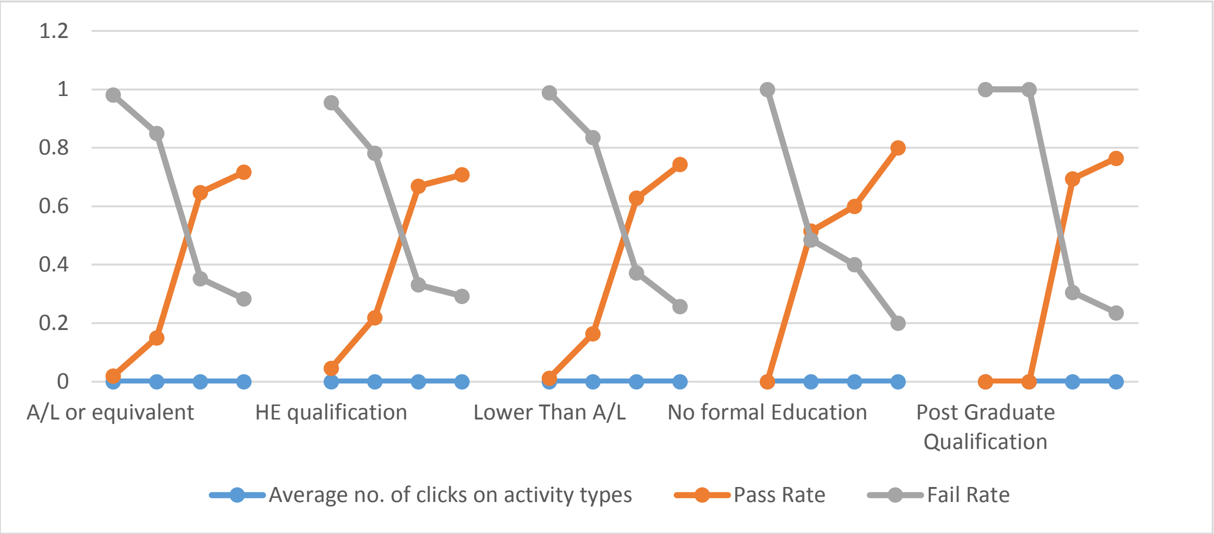


Figure 4.6 Line Chart of VLE activities with Educational Level against Final Result

Figure 4.6 shows when the average number of clicks on activity type increase the pass rate also has increased and at the same time the fail rate has decreased. This pattern is same to the every educational level. This conveys that other than the educational level, the interaction on VLE activity types also influence to the final result of a student.

In previous analysis it was shown that the number of clicks on different activities plays a major role in deciding the final result of a student. Therefore the interactions on each different activity types is considered next, to get better idea on how these activity types influence to the final result of a student.

4.1.3 Interaction on different VLE Activity types with the final result distribution

There are several activity type are used in VLE environment. Table 4.7 figures out how the percentage on number of clicks on different activity types affected to the final result of a student. The table shows both fail and pass students respectively for clear representation on how both category of students interact with the different activity types.

Table 4.7 Cross Table on Interaction on different activity types with different ranges of clicks of pass and fail students at the end of the module.

Clicks	Quiz		Home_Page		OuContent		Forum		Resources		Subpage	
	Pass	Fail	Pass	Fail	Pass	Fail	Pass	Fail	Pass	Fail	Pass	Fail
<=50	0%	47%	5.40%	44.20%	1%	23%	4.40%	48.40%	97%	96%	8%	53%
51-100	0.10%	4.40%	4%	8%	0.20%	5.10%	21.30%	6.60%	3%	4%	7%	8%
101-800	50.80%	31.70%	69.20%	40.60%	22%	47%	55.80%	36.10%	0%	0%	79.10%	34.30%
>801	49.10%	17.10%	21.30%	7.70%	76.80%	25.90%	18.50%	9.00%	0%	0%	6.90%	3.80%

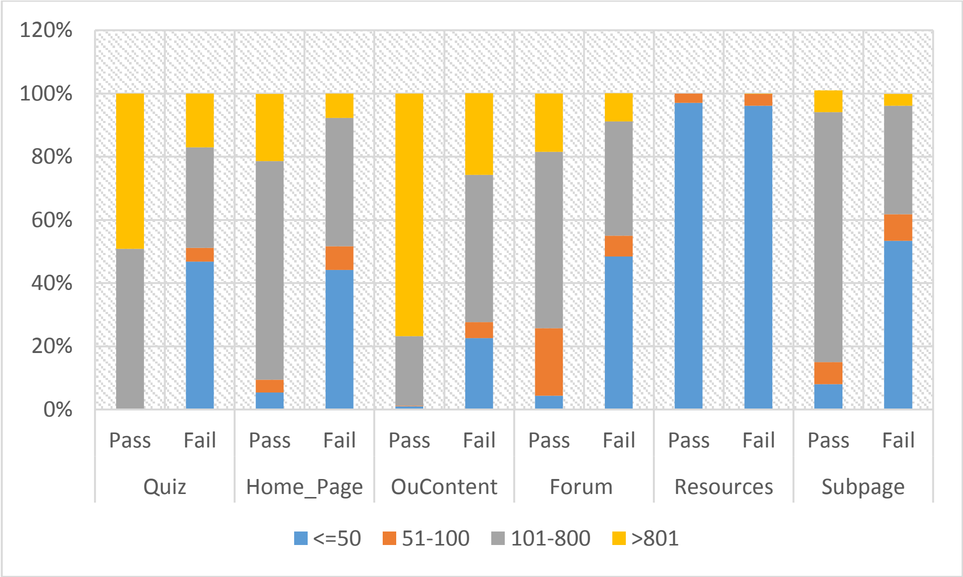


Figure 4.7 Component Bar Chart Table on Interaction on different activity types with different ranges of clicks of pass and fail students at the end of the module.

Figure 4.7 shows a major different with the interaction on different VLE activities by pass and fail students other than Resources. Both pass and fail students have less involvement on Resources where books, journals, articles in pdf format are available for reference. Students who got pass as final result involve with Quizzes mostly with the range of 101-800 numbers of clicks. But this is different with fail students who involve most with lowest range of numbers of clicks. Same differences can be seen in Home Page, Subpages and Forums too. When considering OuConetent which has CMA and TMA assignments, pass students has the highest range on number of clicks on it (>801). For fail students the range on number of clicks (101-800) seems to be high but it is low than the range of pass students. This gives an impression that the reason behind a fail student is because they have less interaction on activity types and also they are unknown on which level they have to face well on these different activity types.

4.1.4 Age Band wise distribution

Students have no age limit to follow the module. Therefore different students with different age distributions are registered to follow the module. The ages are divided into two bands who are below 35 and who are between 35 and 55. Following table 4.6 shows the distribution of the age band on the full dataset.

Table 4.8 Frequency Table for the Age Band wise distribution

Age Band	Rate
<34	73.58%
35-55	26.42%

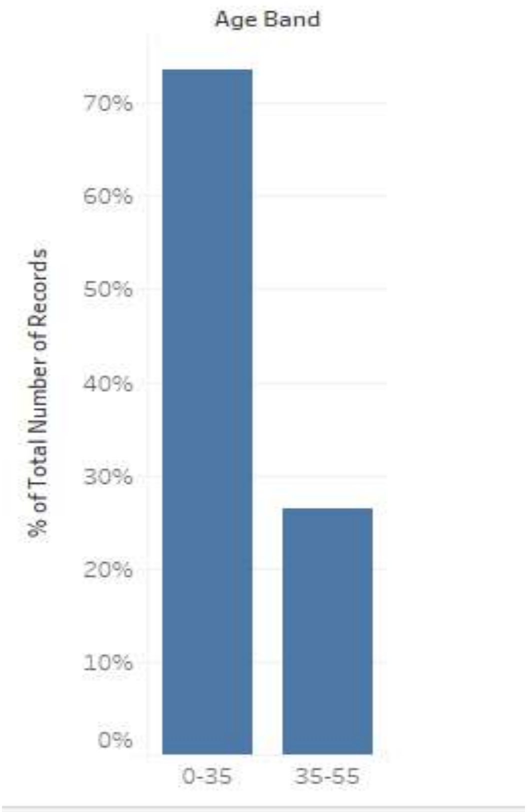


Figure 4.8 Bar chart for the Age Band wise distribution

There are students at young age (below 35) more than 70% who registered to follow this course which is more than thrice of the students who are between 35-55 in their age (only 26.42%)

Next consideration was given to find a relationship between the age band and the pass rate of a student. Table 4.9 is used for this purpose.

Table 4.9 Frequency Table with the final result against age band

Age Band	Pass Rate
0-34	59.22%
35-55	64.35%

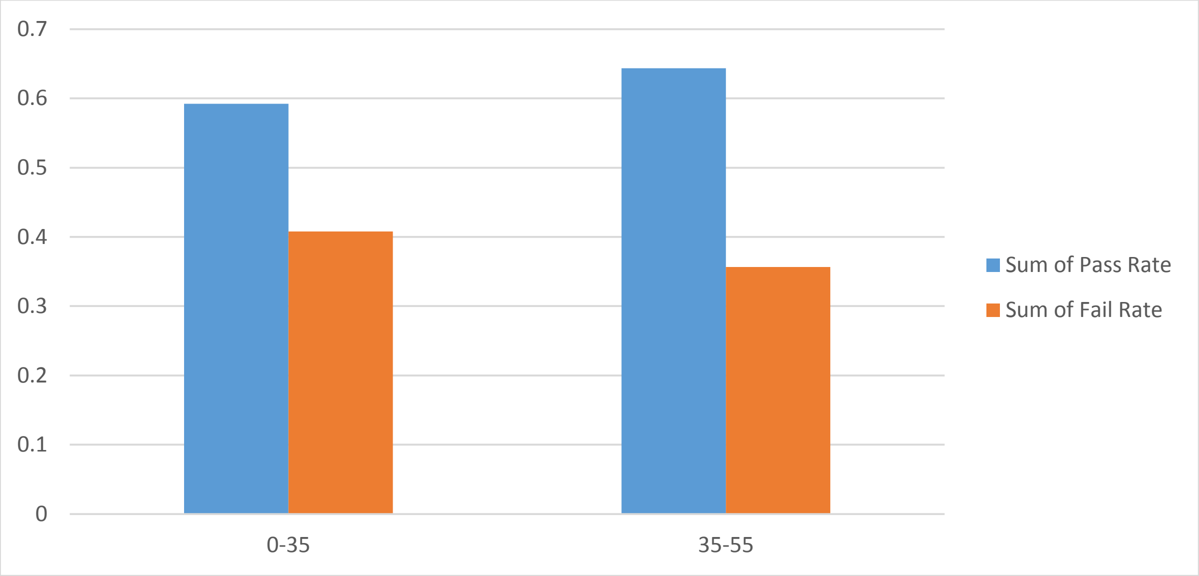


Figure 4.9 Clustered bar chart for the pass rate against age band wise of the distribution

The pass rate is higher of the students who were between 35 and 55 at their age than the students below 35 which is 64.35% and 59.22% respectively. But the considerable amount of fail rate is influenced to the both age groups which is 41% for age category below 35 and 36% age category between 35 and 55.

Next our consideration goes to a main part of a VLE environment which is involvement of different activity types that are available. Following table 4.10 considers whether the involvement of activity types of different age group also help to gain the same result of the above.

Table 4.10 Frequency table for involvement with VLE activity types against age band

Age Band	Percentage on average of click on VLE activity type
0-34	40.92%
35-55	59.08%

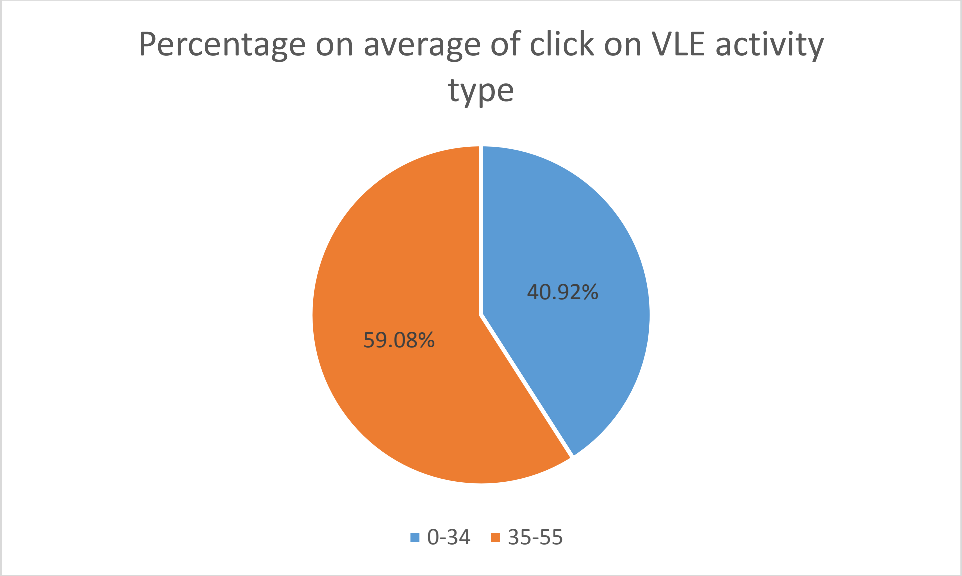


Figure 4.10 Pie Chart for involvement with VLE activity types against age band

Above figure shows that the student between 35 and 55 used VLE activities than the students below 35 in their age which are 59.08% and 40.92% respectively. And also interactions on VLE activity types of students aged 35-55 are more than half when compared to the aged below 35. And also figure 4.9 shows that the students aged 35-55 have higher pass rate / lower fail rate compared to the students aged below 35. This gives an idea that involvement of these activity types is a one factor that affect for the final result of different age groups.

4.1.5 Number of previous attempts wise distribution

There are four attempts allowed to complete the module if the student could not complete the course successfully. In order to find the influence on number of previous attempts on the final result of a student first the consideration was given on how number of previous attempts distributed in the data set.

Table 4.11 Frequency Table with no. of previous attempts wise distribution

No. of previous attempts	Rate
0	85.55%
1	11.47%
2	2.39%
3	0.45%
4	0.13%
Total	100%

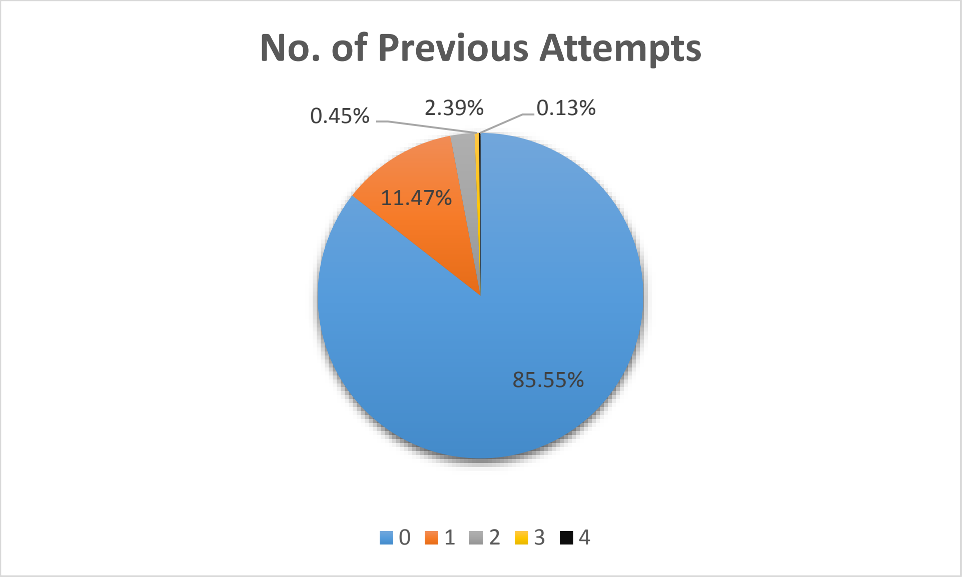


Figure 4.11 Pie Chart with no. of previous attempts wise distribution

85% of students have no previous attempts means that they have attempt to the module at first time. There are only 11% of students with one previous attempt who haven't complete the module previously. The attempts 2, 3 and 4 are not significant when comparing to the attempts 0 and 1.

Table 4.12 shows how previous attempts influence to the final result of the student.

Table 4.12 Frequency Table for pass rate against no. of previous attempts wise distribution

Pass Rate	Attempt 0	Attempt 1	Attempt 2	Attempt 3	Attempt 4
Pass	63.06%	50%	44.06%	37.04%	62.5%

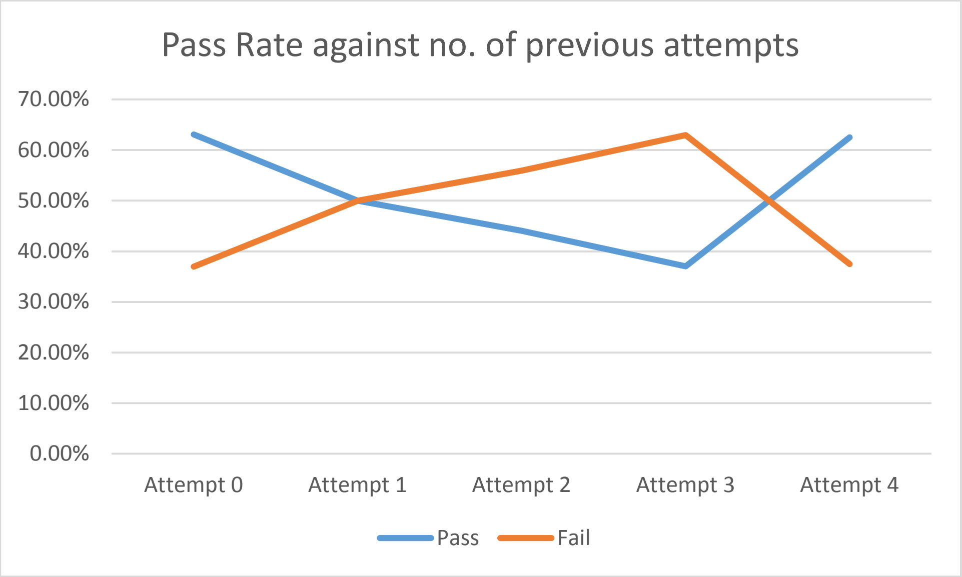


Figure 4.12 Line Chart for the pass rate against number of previous attempts

The pass rate gets decrease and fail rate gets increase when the number of previous attempts get high. But at the 4th attempt which is the last attempt the fail rate is higher than the pass rate. In fourth attempt student has tried on the same module thrice previously without a good completion. This may because the students hasn't identified what is this module about and hasn't much focus on different online activities aided for better final result. In order to pass the final exam even at last attempt students had put their effort on it and get the final result as pass. But the main focus for a success student is to have a good final results without going on several attempts or with less number of attempts. Also the previous analysis show that interactions on different activities is crucial and also the number of interactions on these activities will help for a better final result. Therefore a better guidance need to be given in advance especially for these types of students.

4.1.6 Assignment Score wise distribution

All assignments are compulsory and only TMAs marks are added to the final result. There are five TMA assignments where the weights of those four assignments are 12.5, 12.5, 25, 25 and 25. So all four assignments are considered here where marks given out of 100. These marks were categorized for four ranges and find the distribution of the students on these categories of marks as shown in table 4.13

Table 4.13 Frequency Table for the Assignment Score wise distribution

Assignment Result	Rate
75-100	59.40%
51-74	37.55%
40-50	1.94%
Below 40	1.11%
Total	100%

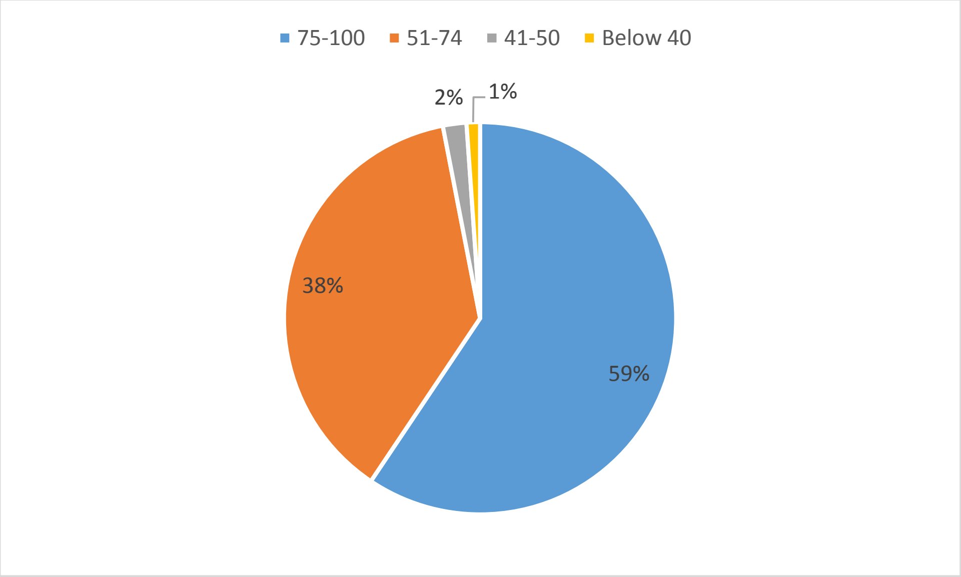


Figure 4.13 Pie Chart for the Assignment Score wise distribution

Figure 4.13 conveys that half of the student in the selected sample has got more than 75 marks for assignments which is in a good condition. One third of students have marks between 51 and 74. 41-50 marks and the fail (below 40 marks) students are less compared to the other marks.

As the marks of these types of assignments are added for the final result it is better to consider on following frequency table 4.14 that shows how assignment scores influence to the final pass rate by considering the whole data set.

Table 4.14 Frequency Table for the pass rate against assignment score wise distribution

Assignment Score	75-100	51-74	50-40	Below 40
Pass	45.25%	15.77%	0.02%	-
Fail Rate	14.19%	21.79%	1.92%	10%

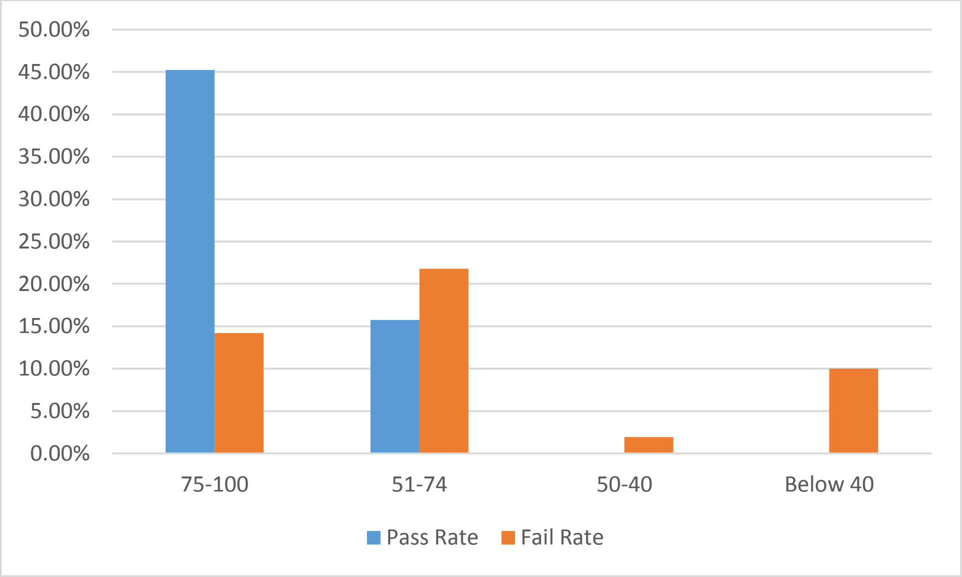


Figure 4.14 Clustered bar chart for pass rate against Assignment score wise distribution

From the entire sample highest pass rate which is nearly half was taken by students who get more than 75 marks for the assignment. Students who get marks between 51 and 74 for the assignment have nearly 16% pass rate and 22% fail rate which seems failure rate is high. Also Students who have marks between 40 and 50 for the assignment have higher failure rate than pass rate. All students who below 40 marks for the assignment have the final result as fail. This seems if a student get a mark for the assignments 75 out of 100 then he/she has higher probability to gain a pass as final result of the module when compare with other students who have below 75 marks for assignments.

4.1.7 Conclusion of Initial Descriptive Analysis

The initial descriptive analysis was done to get familiarized with the data set and through that get an in-depth knowledge on each individual attribute in the data set. Also multiple attributes analyzed together especially with final marks and number of clicks on different activity types. Using graphical techniques the results of the descriptive analysis was presented to the reader.

Finally the objective of this research is to monitoring the performance of a student in an e-learning environment by collecting some attributes from a student demographical and e-learning environment, complex relationships and patterns should be considered and identified between the final result and the other attributes. Analyzing variables one by one using statistical method as was done in initial descriptive statistics would be a very complex approach for complex pattern identification and also would be a very time consuming process. Most importantly the accuracy of such a model and its' result is questionable. Hence an advanced analysis using data mining algorithm has to be used for rest of the study in order to find complex patterns between the final results and other attributes.

4.2 Advanced Analysis using Learning Algorithm

After obtaining a good understanding about the dataset, data mining concepts were used in Multilayer Perception which is a Neural Network Algorithm in order to identify most affected attributes for the final result, Decision Tree which shows the relationship between affected attributes and Regression Model which estimates the probability of pass/fail of a student. For the purpose of model evaluation and accuracy evaluation, Confusion Matrix, ROC curves and Gain Charts are used.

Table 4.15 shows how data set is arranged in order to apply the data mining techniques. Among the respondent 70% of data was used to train the model and 30% of data was used to test the accuracy of the model.

Table 4.15 Arrangement of dataset

		Number of students	Percent
Sample	Training	4190	69.6%
	Testing	1832	30.4%
Valid		6022	100.0%

4.2.1 Neural Network Model

During the advanced analysis done for the students the variables used are Gender, Age, Disability, Highest Education, No. of previous attempts, Average Score for assignments and No. of clicks for Forums, Home Page, OUcontent, Quiz, Resources and Subpage.

For the above sample set Neural Network was applied. In the procedure, minimum number of hidden layers that can have for the model was given as 1 and batch wise training was used.

Table 4.15 depicts how the factors were given for the model. Altogether there were 12 factors considered. Final Result (pass/fail) has been used as dependent variable which will be predicted by the model. Hyperbolic Tangent method was used as the activation function for the evaluation of the hidden layers and Softmax method was used for the analysis of the output layer.

Table 4.16 Neural Network Parameters Used

Input Layer	Covariates	1	Gender
		2	Age_Band
		3	Disability
		4	Highest_Education
		5	No_Of_Prev_Attempts
		6	Avg._Score
		7	Forum
		8	Home_Page
		9	Oucontent
		10	Quiz
		11	Resources
		12	subpage
	Number of Units[a]		12
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		8
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Final_Result
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy
a. Excluding the bias unit			

Standardized: Subtract the mean and divided by the standard deviation. , (x-means)/s.

Hyperbolic tangent: This function has the form: γ(c) = tanh(c)=(e c −e −c)/(e c +e −c). It takes real-valued arguments and transforms them to the range (–1, 1). When automatic architecture selection is used, this is the activation function for all units in the hidden layers

Softmax. This function has the form: $\gamma(c\ k) = \exp(c\ k)/\Sigma j \exp(c\ j\ )$. It takes a vector of real-valued arguments and transforms it to a vector whose elements fall in the range (0, 1) and sum to 1. Softmax is available only if all dependent variables are categorical. When automatic architecture selection is used, this is the activation function for units in the output layer if all dependent variables are categorical.

Table 4.17 Neural Network Model Summary with Error Rates

Training	Cross Entropy Error	1197.862
	Percent Incorrect Predictions	11.7%
Testing	Cross Entropy Error	568.589
	Percent Incorrect Predictions	13.4%
Dependent Variable: Final_Result		
Error computations are based on the testing sample.		

Table 4.17 displays a summary of neural network results by partition and overall, including the error and percentage of incorrect predictions. The error (cross-entropy error) when the softmax activation function is applied to the output layer. It can be seen that in the Training dataset the incorrect prediction percentage is 11.7% which is acceptable in a practical data set. It can also see that in the testing sample the percentage of incorrect prediction is also 13.4%. This shows that the weights assigned to the model are in an acceptable level.

4.2.1.1 Neural Network Experiment Results

Table 4.18 performs a sensitivity analysis which computes the importance of each predictor. The analysis is based on the combined training and testing samples. The value 'Normalized importance' is simply the importance value divided by the largest importance value and expressed as a percentage. Figure 4.15 shows the graphical representation of the table 4.16

Table 4.18 Importance of different factors affecting for the Final Result

	Importance	Normalized Importance
Gender	.013	8.3%
Age_Band	.006	4.1%
Disability	.021	13.8%
Highest_Education	.027	17.9%
No_Of_Prev_Attempts	.056	36.8%
Avg._Score	.141	92.5%
Forum	.136	89.6%
Home_Page	.070	46.0%
Oucontent	.152	99.7%
Quiz	.152	100.0%
Resources	.077	50.3%
subpage	.149	98.2%

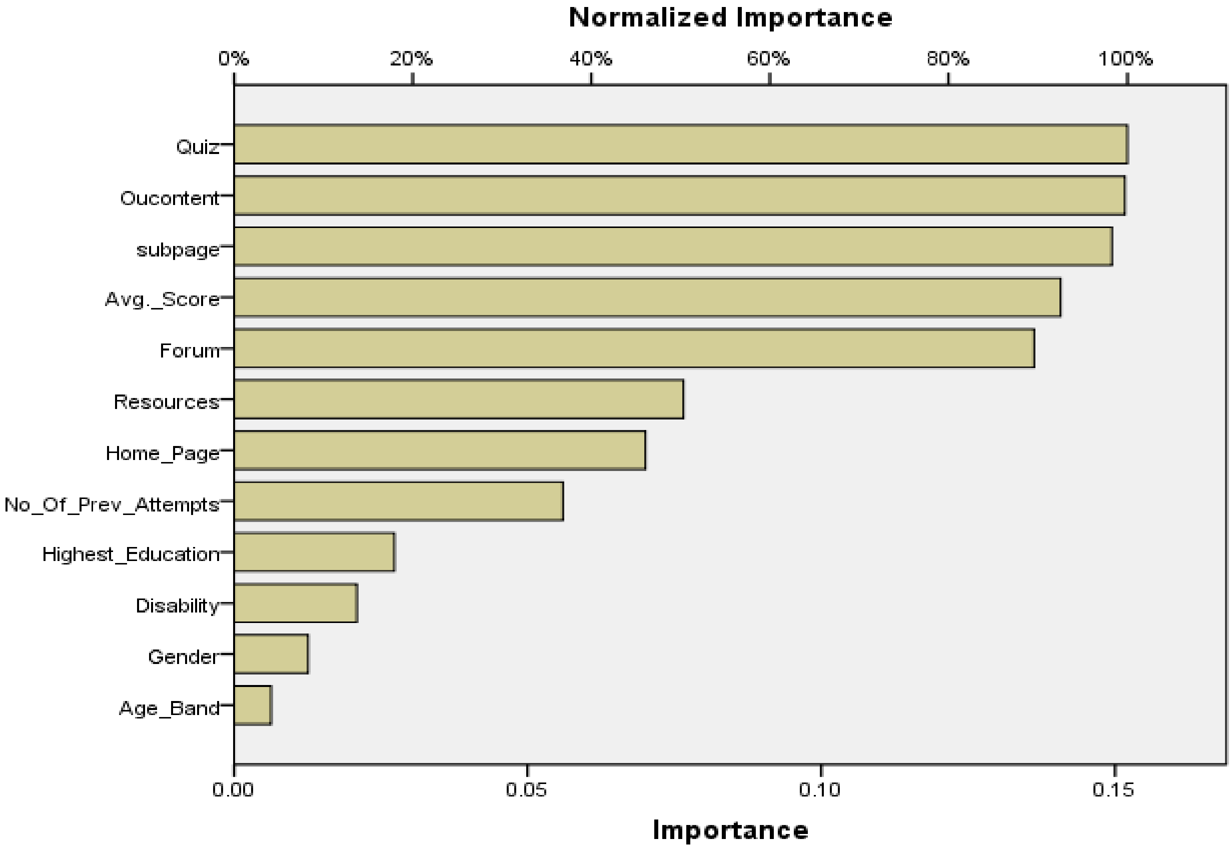


Figure 4.15 Importance of the different factors affecting for the final result

According to the table 4.18 and figure 4.15 given by the neural network model, it can be seen that number of clicks for the quizzes play the most important role in deciding the final result of a student although Quizzes have no any marks that added for the final results. The probability of importance of the no. of clicks on Quizzes is 0.152. Second highest importance which is not much vary with the weights of Quiz is the OUContent which represents contents of assignments which should pass with computer marked assessments where the marks are not added to the final result. Next importance is that which represents the subpages that point to other sites that relevant to the course together with basic instructions. Average marks of assessments where the marks are added to the final result also play important roles in this model but quizzes, oucontents and subpages are more important than assignments.

Resources which usually contains pdf resources like books, Home Page where notices and updates on the course are available, forum discussions give a significant influence on the final result of this course. As the learning materials and methods of this course is mainly depend on an e-learning environment it can be seen that the functions used in an e-learning are take major roles than demographic attributes of a student such as Number of previous attempts, Highest Education, Disability, Age and Gender.

Also it can be noticed that among e-learning methods that mostly effect with marks like quizzes, OUContent, subpages have become more important than others and resources have the least important.

This gives the impression that the demographical factors will not influence to pass a course if it used an e-learning environment but a particular student should focus on the various e-learning methods used for the course.

4.2.1.2 Neural Network Model Evaluation

Classification Table

Table 4.19 Classification Table for Neural Network Model

Sample	Observed	Predicted		
		Fail	Pass	Percent Correct
Testing	Fail	536	182	74.7%
	Pass	63	1051	94.3%
	Overall Percent	32.7%	67.3%	86.6%
Dependent Variable: Final_Result				

The classification table shown in table 4.17 provides a summary of predicted values against the actual values. Actual values are named as observed values. For a given case suppose the observed value and the predicted value is same. Then it is said to be True Positive (TP) if it is observed and predicted as ‘Passed’ or ‘True Negative’ (TN) if it is observed and predicted as ‘Filed’. Cells in the primary diagonal of the table illustrate the correct prediction.

The classification table given in 4.15 is a summary for testing and training samples. Accuracies of the predictions are given in three ways. They are the accuracy of prediction failures, prediction passes and the accuracy of overall prediction (both passes and failures).

Table 4.15 shows 182 cases which were incorrectly classified as pass which are actually fail and also 63 cases which were incorrectly classified as pass although they were fail. There is a significant effect if a pass student is incorrectly classified as a fail student than a fail student is incorrectly classified as a pass student.

From the table it can be seen that for the testing set the accuracy is 86.6% and provides evidence that the weights calculated by the Multilayer Perception procedure seems to be accurate.

Cumulative Gain Chart and Lift Chart

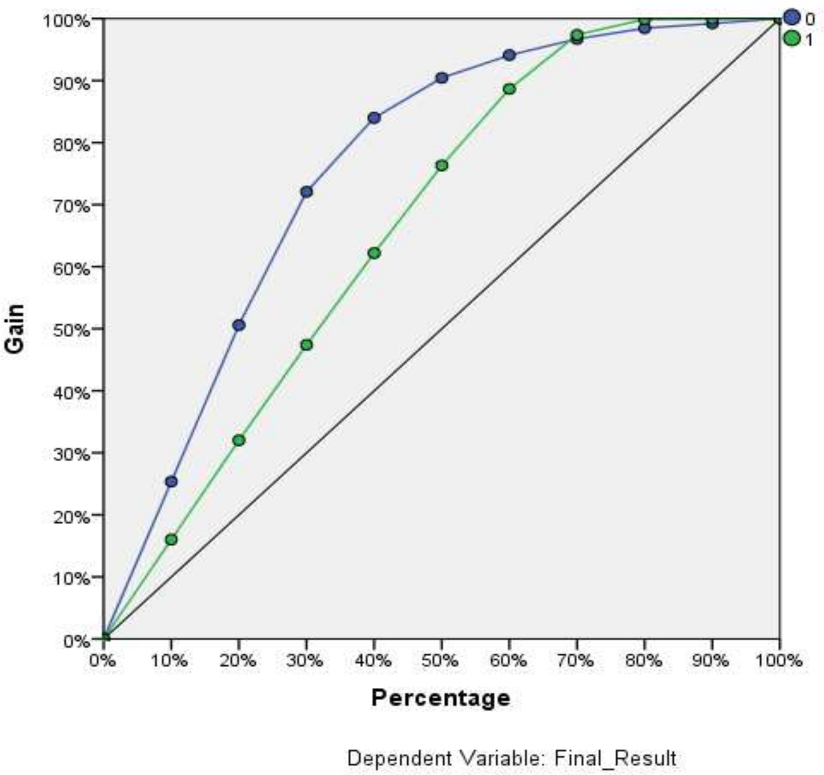


Figure 4.16 Cumulative Gain Chart for Neural Network Model

The Cumulative Gain Chart in figure 4.16 shows the percentage of the overall number of cases in a given category 'gained' by targeting a percentage of the total number of cases. For example, the first point of the curve for the 'Pass' category is at (10%,17%) meaning that if a dataset is scored with the network and sort all of the cases by predicted probability of 'Pass", you would expect the top 10% to contain approximately 17% of all the cases that actually take the category 'Pass' (defaulters). Likewise the top 70% would contain approximately 90% of the defaulters. 100% of the dataset obtain all of the defaulters.

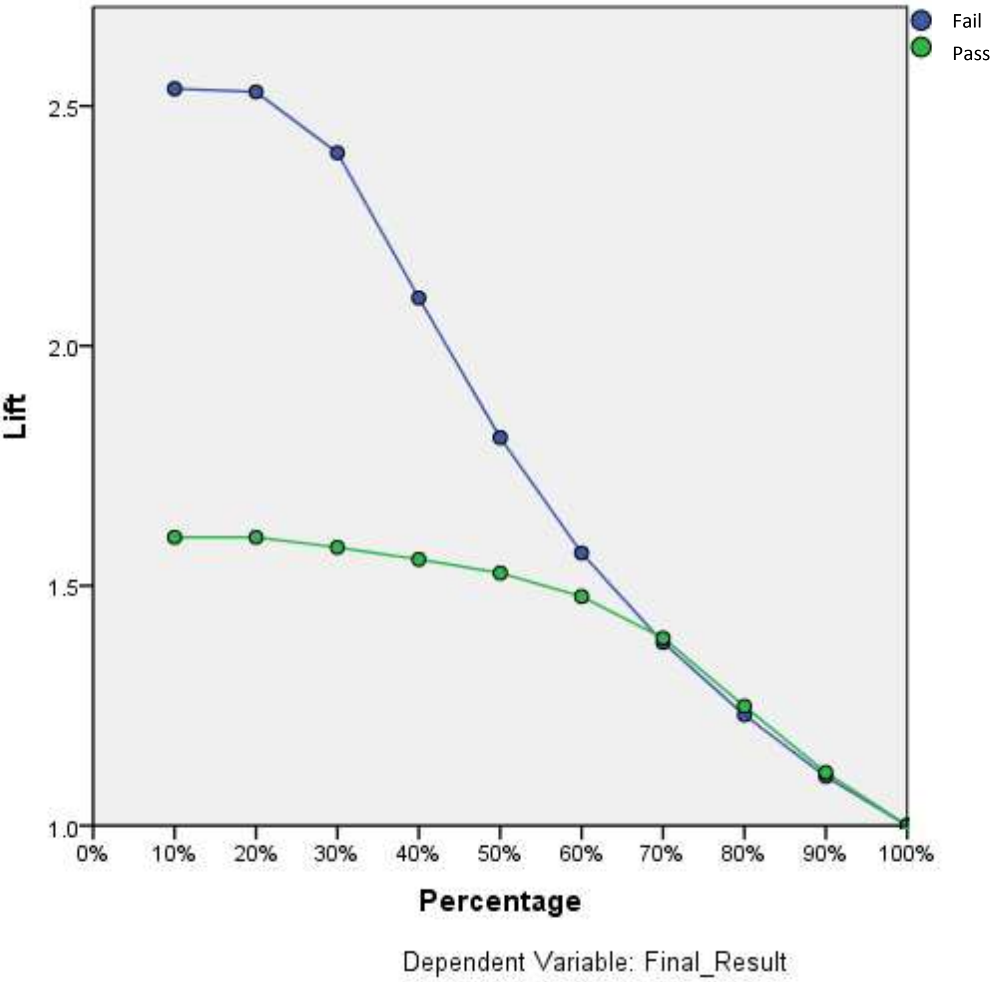


Figure 4.17 Lift Chart for Neural Network Model

This measures how much better one can expect to do with the predictive model comparing without a model. The Lift Chart is derived by the Gain Chart. For example the lift at 10% for the category 'Pass' is 17% divided by 10% which is 1.7. This means that when selecting 10% of the records based on the model, one can expect 1.7 times the total number of targets (students who got pass as final result) found by randomly selecting 10% without a model.

4.2.2 Decision Tree

Greatest effects on how the network classifies students on their final result was identified. But the direction of these relationships between effected factors are unable to identify using MLP. CHAID is a tool used to discover the relationship between variables. This analysis builds a predictive model or tree to help determine how variables best merge to explain the dependent variable final result.

Same data sample was used for training and testing where the sample sets are depicted in table 4.15

Table 4.20 depicts how the factors were given for the model. Altogether there were 12 factors considered. Final Result (pass/fail) has been used as dependent variable which will be predicted by the model.

Table 4.20 Decision Tree Parameters Used

Specifications	Growing Method	CHAID
	Dependent Variable	FinalResult
	Independent Variables	Gender, AgeBand, Disability, HighestEducation, NumOfPrevAttempts, Avg.Score, Forum, HomePage, Oucontent, Quiz, Resources, subpage
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	200
	Minimum Cases in Child Node	100
Results	Independent Variables Included	Quiz, Avg.Score, Oucontent, Resources, subpage
	Number of Nodes	23
	Number of Terminal Nodes	15
	Depth	3

Table 4.20 shows some very broad information about the specifications used to build the model and the resulting model. The specifications section provides information on the settings used to generate the tree model including the 12 factors. The result section displays information on the number of total and terminal nodes, depth of the tree (number of levels below the root node) and independent variables included in the final model. 12 independent variables were specified but only five were included in the final model. The variables Gender, Age Band, Disability, Highest Education, Number of previous attempts, Home Page and Forum did not make a significant contribution to the model, so they were dropped from the final model.

4.2.2.1 Decision Tree Experiment Results

In practice CHAID often used to understand how different group of students respond to the final result on their behavior on e-learning environment. Figure 4.18 shows the direction of effect of different factors on the final result of the student.

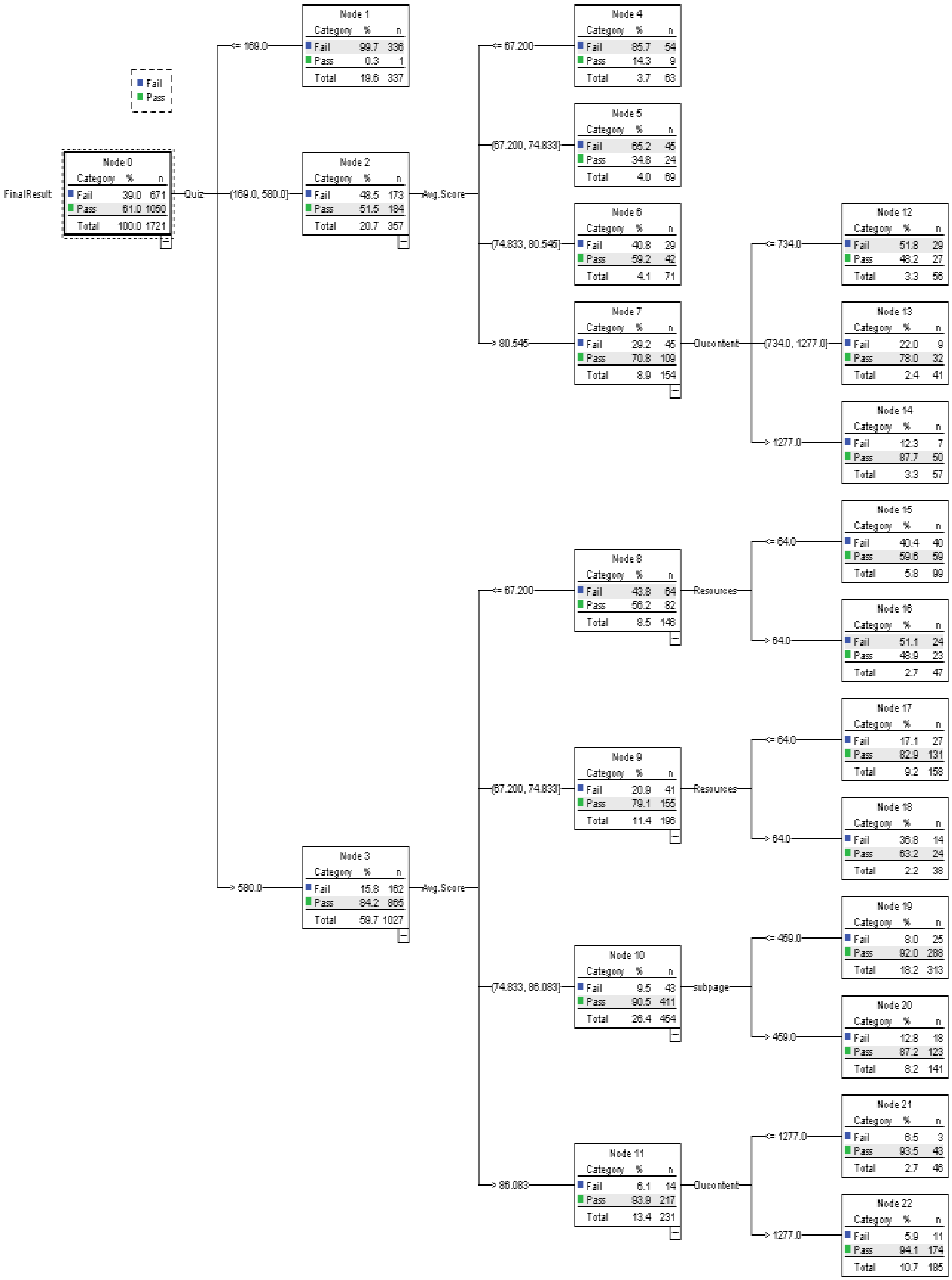


Figure 4.18 CHAID Tree Diagram showing the factors for the final result of a student

At the first level (trunk) we have all students with 60.9% of pass students and else fail students. As the tree is progressed down to the first branch, the factor that has the greatest impact on the likelihood of final is identified which is Quiz. Then the overall sample is broken down in to groups (leaves) based upon their differing values on number of interactions on Quizzes. (169 or less, higher than 169 or lower than or equal 580 and higher than 580). For continuous variables CHAID use discretization to make it as categorical. These categories have pass rate of 0.3, 51.5 and 84.2 percentages respectively. So interaction with Quizzes get high then the pass rate also gets high. Then the difference is checked for statistically significance and if it is those are retain as new leaves. Then it is considered whether leaves can be further split to subgroups. Students who have interaction with more than 580 times on quizzes and who have more than 86 of average mark for the assignments have higher pass rate of 93.9% when compare with the other categories of students. At each step every predictor variable is considered to see if splitting a sample based on a factor leads to a statistically significant relationship with the response variable final result.

Using terminal nodes of the decision tree rules can be generated as a set of logical "if-then" statements that describe the model's classification or predictions for each node. Rules with highest probability on pass and fail are stated below while other rules are on Appendix B.

Rule for Pass condition

If (Quiz >580 AND Average Score >86.08 AND Oucontent > 1277) then Prediction =Pass where node=22 and probability = 0.974

This shows that if a student has interactions on Quizzes and Oucontent in higher rates and also has average score more than 86.08 then that student has the highest probability to have the final result as Pass which is 0.974

Rule for Fail condition

If (Quiz <=169) then Prediction = Fail where Node=1 and probability=1

This shows if a student has less interaction (less than 169) on Quizzes then that student has the highest probability of 1 for the final result as Fail. And also it can be noticed that there no any other interactions are considered for this decision.

4.2.2.2 Decision Tree Model Evaluation

Risk Table

Table 4.21 shows the risk table of the decision tree which provides a quick evaluation of how well the model works.

Table 4.21 Risk Table for the CHAID Decision Tree model

Estimate	Std. Error
.155	.009

Test sample results are displayed.

The risk estimate of 0.155 indicates that the category predicted by the model (pass/fail) is wrong for 15.5% of the cases. So the risk of misclassifying a student final result is approximately 16% which is acceptable.

Classification Table

Table 4.22 Classification Table for the model CHAID Decision Tree

Sample	Observed	Predicted		
		Fail	Pass	Percent Correct
Testing	Fail	488	183	72.7%
	Pass	84	966	92.0%
	Overall Percentage	33.2%	66.8%	84.5%

This table shows the number of cases classified correctly and incorrectly for each category of dependent variable. Thus the model classifies approximately 84.5% of the students correctly. Also 72.7% of fail students and 92% of pass student classifies correctly by CHAID Decision Tree model.

4.2.3 Regression Model

Using NN highest influence factors on Final Result of a student identified. Decision Tree shows relationships among the influenced factors. If one needs to build up a system it would be better to derive an equation with all possible factors influenced for a student’s final result. Therefore the analysis is next concentrated on deriving a regression equation for the predicted variable. Since the final result is dichotomous variable, Binary Logistic Regression model is used. Same data used as in Table 4.15 for testing and training purpose.

Table 4.23 gives the overall test for the model that includes the predictors. The chi-square value of 770.543 with a p-value of less than .0005 tells us that our model as a whole fits significantly better than an empty model (that is a model with no predictors) at 95% level of confidence. (Reject null hypothesis). df represents the degree of freedom for each predictor in the model.

Table 4.23 Coefficient test of Binary Logistic Model

	Chi-square	df	Sig.
Model	770.543	12	.000

Table 4.24 represents the accuracy of the model. The -2 Log likelihood (1304.787) can be used in comparison of nested models. The Negelkerke R^2 and Cox and Snell R^2 which refers to as pseudo R^2 which are interpret in the same manner. Therefore the explained variation (total variance -residual variance) in the dependent variable based on our model ranges from 30.4% to 48.6% depending on whether the use of two R^2 s. Anyway the Cox and Snell R^2 cannot achieve the value of 1. For this reason as it is preferable to use Negelkerke R^2. Therefore the value 0.481 is a decent value for a practical dataset like this.

Table 4.24 Model Summary for Binary Logistic Regression

-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1304.787	.304	.486

4.2.3.1 Binary Logistic Experiment Results

Table 4.25 shows the contribution of each independent variable to the model and its statistical significance.

Table 4.25 Variables in Binary Logistic Model

	B	S.E.	df	Sig.	Exp(B)
Gender	.149	.190	1	.432	1.161
Age_Band	-.007	.183	1	.971	.993
Disability	-.783	.212	1	.000	.457
Highest_Education	.188	.105	1	.073	1.207
No_Of_Prev_Attempts	-.274	.116	1	.018	.760
Avg._Score	.105	.011	1	.000	1.111
Forum	.000	.000	1	.098	1.000
Home_Page	.000	.001	1	.430	1.000
Oucontent	.001	.000	1	.003	1.001
Quiz	.007	.000	1	.000	1.007
Resources	-.004	.002	1	.055	.996
Subpage	-.005	.001	1	.000	.995
Constant	-11.241	.902	1	.000	.000

The statistical significance of the test is found in the Sig. column. From above table it is shown that Disability (sig.<00005), Number of previous attempts (sig.=0.018), Average score (sig.<0.0005), Oucontent (sig.=0.003), Quiz(sig.<0.0005), subpage (sig.<0.0005), constant (sig.<0.0005) added significantly to the model as the sig. value is less than critical value which is 0.05

But, Gender (sig.=0.432), Age Band (sig.=0.971), Highest Education (sig.=0.073), Forum (sig.=0.098), Home Page (sig.=0.430) and Resources (sig.=0.055) did not add significantly to the model as sig. value is greater than the critical value 0.05.

The Exp(B) shows odd ratios and indicates that number of intersections on OuContent and Quizzes and the average mark for the assignment increased by one unit it will influence one time as likely to the final result.

The B column shows the coefficient values for each variable which is used when generating the equation. Using the table 4.22 the logic model can be derived as follows.

$$\ln(P) = -0.783X_1 - 0.274X_2 + 0.105X_3 + 0.001X_4 + 0.007X_3 – 0.005X_4 - 11.241$$

Where, X_1 = 1 if a student is disabled

X_2 = 0 if there is no any number of previous attempts

X_2 = 1 if there is a one previous attempt

X_2 = 2 if there are two previous attempts

X_2 = 3 if there are three previous attempts

X_2 = 4 if there are four previous attempts

X_3 = Average marks for the assignments

X_4 = Number of clicks made on Oucontent

X_5 = Number of clicks made on Quizzes

X_6 = Number of clicks made on subpages.

By the logit equation we can gain estimated probability (p) by

$$p = \frac{1}{1+e^{-logit(P)}}$$

4.2.3.2 Binary Logistic Model Evaluation

Classification Table

Table 4.26 Classification Table for Binary Logistic Model

Sample	Observed		Predicted		
			Final_Result		Percentage Correct
			Pass	Fail	
Training	Final_Result	Pass	1015	118	89.6%
		Fail	204	468	69.6%
			67.5%	32.5%	84.7%

Table 4.26 shows the number of cases classified correctly and incorrectly for each category of dependent variable. Thus the model classifies approximately 89.6% of the students correctly. Also 69.6% of fail students and 84.7% of pass student classifies correctly by the Binary Logistic model.

4.2.4 Comparison of Models

Accuracy of the classification table

By using the values of classification tables from table 4.16, table 4.19 and table 4.23 the prediction accuracies are calculated as shown below for neural network model, decision tree model and Regression model. The method of calculating the prediction accuracy is explained in detail in the Methodology Chapter.

$$\text{Accuracy} = \frac{TP+TN}{N}$$

Here by applying this method to training data in NN MLP model the accuracy of the model can be calculated as 86.6%.

By applying this method to training data set in CHAID decision tree model the accuracy of the model can be calculated as 84.5%.

By applying this method to training data set in Binary Logistic Regression model the accuracy of the model can be calculated as 84.7%.

It can be seen that when classification tables are used for model evaluation the above models perform well.

Area under Curve

The ROC curve shown in figure 4.18 provides a visual display of the Sensitivity and Specificity for all used models in a single plot, which is much cleaner and more understandable than a classification table. The chart displays three curves respectively for the Multilayer Perception Model (MLP), CHAID Decision Tree model (CHAID) and Binary Logistic Regression Model (BLR).

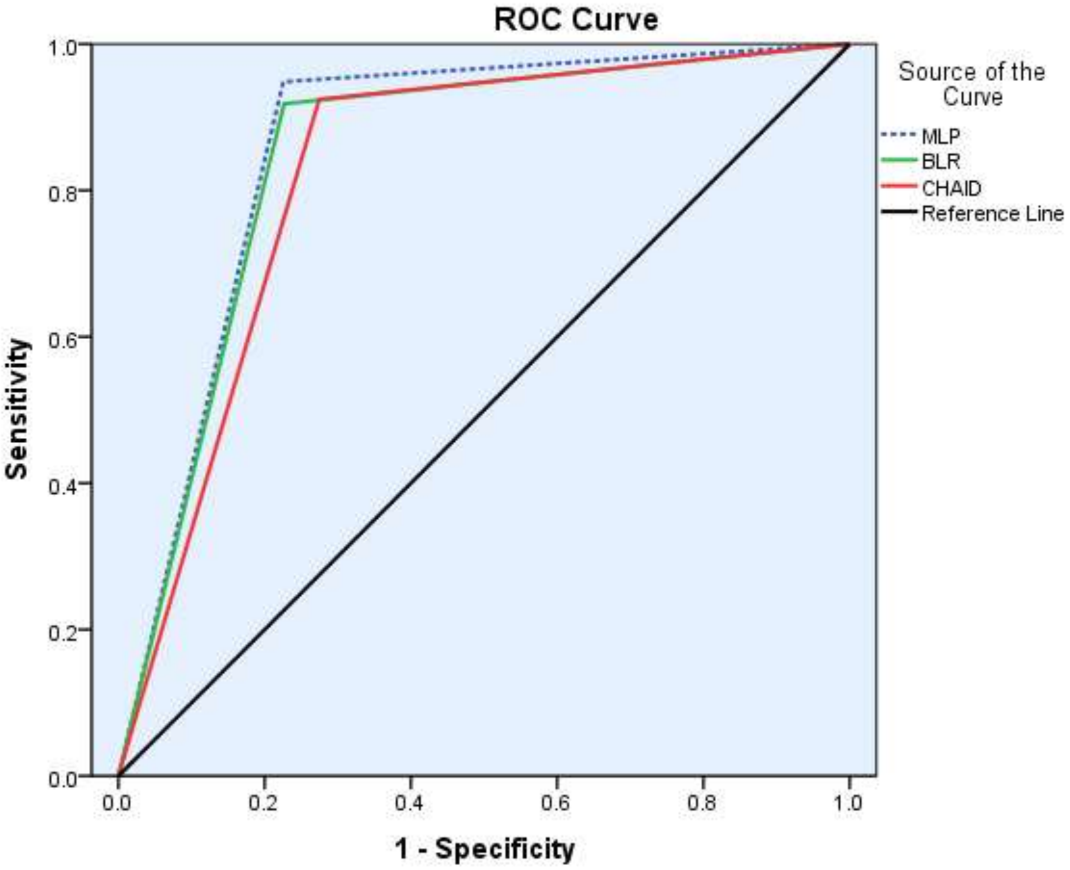


Figure 4.19 ROC Curve for MLP, CHAID and BLR models

When considering the curves, Sensitivity refers to the True Positive (TP) rate (Y axis), the probability of correctly diagnosing a positive case (Pass as Pass). Specificity refers to False Positive (FP) rate (X axis), the probability of incorrectly diagnosing a case as positive when its true state is negative (Fail as Pass). The curve climbs quickly towards the top left meaning the model correctly identifies cases and then gradually moves towards toward top right corner. And also the curve doesn’t come to the 45 degree diagonal of the ROC space indicates that the model has a high accuracy. The reference line in the figure refers to a prediction being correct when a data point is picked arbitrary.

The area under curve shown in the table 4.27 is a numerical summary of the ROC curve. The Area Under Curve (AUC) is a measure of the ability to diagnostic test to correctly identify cases. Diagnostic tests with higher AUCs are generally better and should always be higher than 0.5, indicating better the test id diagnosing than an arbitrary prediction.

Table 4.27 Area Under the curve for ROC

Model	Area Under Curve
Multilayer Perception Neural Network	0.862
CHAID Decision Tree	0.825
Binary Logistic Regression Model	0.846

ROC curve analysis it is shown that Multilayer Perception model has the highest AUC when compare with the other models that can be concluded it has a slightly high sensitivity than the CHAID Decision Tree and Binary Logistic Method. In order to identify important factors affecting to the final result of a student Multilayer Perception method can be used. To show the direction of the affecting factors of the final result the CHAID Decision Tree can be used and Binary Logistic Regression Model can calculate an exact value given by an equation.

4.3 Summary

Firstly an initial descriptive analysis was carried out in order to understand the data set. Attributes of the data set was analyzed individually and then relationship among few attributes with the Final Result were analyzed using statistical methods and the results were discussed.

After the initial analysis advanced analysis was used to find out hidden relationships between the attributes and the Final Result. In order to analyze the effects of these factors on the Final Result the data were analyzed using Neural Network Multilayer Perception method, CHAID Decision Tree and Binary Logistic Regression Model. Results obtained by these methods were discussed in this chapter and finally a discussion is presented evaluating and comparing the accuracies of the models.

An overall discussion on all the results obtained in this study is presented in the following chapter Conclusion. The limitation faced in the study and future work in this research are also discussed in the next chapter.

Chapter 5: Conclusion

5.1 Discussion and Conclusion

During this research it was targeted to find how several methods used to perform the educational modules in an e-learning environment affect the student performance to pursue for an excellent result in order to finish the module in an success way. Also these findings, give a better idea for a tutor for motivating their students on the way of using these important methods by monitoring them in an e-learning environment and to design their course modules by giving priority for those important methods.

The study was conducted in many stages. In the first stage the consideration was given to understand the problem. Then a suitable data sample was collected. Once the dataset was collected it was preprocessed and cleaned due to various issues identified in the dataset. A descriptive analysis was conducted to get familiarized with the dataset. Main factors used for analysis were number of clicks on forum discussions, home pages, Oucontent which represents content of assignments which should pass, quizzes, resources contains pdf resources as books, subpage that points to other sites in the course together with basic instructions which are recorded by Moodle itself and demographical data like age, gender, disability, highest educational qualification and number of previous attempts for a particular course.

Mainly three data mining techniques were used in the study in the advanced analysis process. Those are Neural Network as Multilayer Perception, Decision Tree as CHAID and Regression as Binary Logistic.

NN model used to find most affected factors for the final result. It was noticed that according to NN models, Quizzes, attempt in the oucontent which include assignments, subpages that point to the other sites of the course and marks for the assignments are the factors which has the highest influence on the final result.

On the other hand online quizzes give major influence for the final result. By using quizzes student can have an immediate reaction for the answers. As teacher is not directly involved in e-learning environment this gives opportunity for both parties as student can read teacher-provided strategy for improvement for each wrong answer. Students do not just know that they are incorrect but they see an explanation on how to improve. Both students and teachers can see the students' progress over the time as they can see the online quizzes scores at that time. Instructors then can determine how to help for students on their weak activities. Students can know easily in what section that he/she must consider most. Therefore it is no doubt that quizzes can help to enhance the performance level of a student.

Oucontent have assignments that should pass. Altogether there are twelve assignments with seven CMA assignments and five TMA assignments. Although CMA assignments give marks out of 100 those marks are not taken for calculation of final result. CMA assignments are made using with one block of module or several blocks of the module together. There is no deadline for a CMA assignment and student can face for the questions at any free time. Purpose of these types of assessments is to identify strength and weaknesses in one's own work and revise accordingly. Students can gather feedback for themselves in order to make improvements before it is due for grading.

Subpages give an interactive environment for students who enjoy with different types of learning styles via other links. Student can get an idea on how the learning theories have been adapted in practical environment through this. Students who cannot understand the written theories can join with these types of sources for verbal and visual learning methods.

TMA assignment marks are added for the final result and the feedback of those assignments are also provided. Students have to submit the assignments on due date. Average score out of 100 were considered here where the weighted scores for five assignments are 12.5,12.5,25,25 and 25 respectively.

Another important factor identified by the NN was that the methods used in e-learning environment is more important than demographical data of a student. Among those number of previous attempts for the course come first. It shows that most of students who register for e-learning courses have different demographics and none of them will influence for the performance of a student.

The decision tree model also reveals that in order to get better results student should mainly focus on the Quizzes and assignments in oucontent. It seems both MLP model and CHAID model shows the highest influence is on formative assignments where marks are considered as self-evaluation criteria. CHAID decision trees shows how the number of clicks affected and the relationship among those different factors. It shows if a student have less than166 number of clicks on quizzes then he has higher probability to fail in the final exam. This will help for the student to check his progress before the final examination and also for the instructor to motivate the student for a better result by monitoring the student performance mainly on these fields. But decision tree gets complex if more factors are used to build it.

Binary Logistic Regression Model gives an opportunity to find the probability of the pass of a student with varieties of factors. Student can be evaluated whether he is in near the pass result or whether he has to focus on more. BLR shows that gender, age, highest education, forum, home page and resources are not statistically significant for a performance of a student. Quizzes, Oucontent and Average Score are statistically significant in this modules and also those are the mostly affected factors in MLP and CHAID decision tree. Other than those, Disability and Number of previous attempts also influence significant but it is negatively for the final result.

Although there are many benefits of e-books the studies have shown that the use of e-books give poor feedback on progress of a student. Generally the apps for e-reading lack the ability to present essential landmarks and make it difficult to plan readings. This has less interactive environment as it is only a direct transfer of a hard copy. Also most of the students who learn using an e-learning system are students that involve with other activities in their personal life which limits them the time to learn. So they have no time to read additional books. Also Forums have negative impact on students' performance. This may because the instructors are not give priority for the forum discussions by involve themselves with this. Also as there are more than thousands number of students studying this course, and if all students involve with the discussions, others have to give more time to read others ideas on a subject matter.

Using this knowledge and by identifying the real reasons why these factors are very important, necessary guidance and support could be given to students providing ways and means to improve these factors and knowledge. The study gives hints to the instructors to format the course contents by using different kinds of e-learning methods in an appropriate way from matching the students' learning style to get best performance of learning process. Students can be advised in advance that he is walking on a wrong path for a successful end and what flowers to pick out from the garden with variety of flowers without making him lose.

The result of this study can be used for any organization that use e-learning as a method of learning to plan the program accordingly, setting standards and in turn produce students with great performance by monitoring the student performance.

5.2 Future Work

The research results presented in this thesis could be used to develop an innovative strategy to make modern planning and developing up to date learning resources content using innovative technology, implementing the new learning content.

The study could be applied with other courses, modules and lecture for the academic team in general, and a large sample of students in coming academic years. A new academic approach for performance Also this research can be extended to other learning softwares other than Moodle. The selected factors can be further analyzed in time series to see which attributes that use in each factors affected really for the students final result other than taking the factor as a whole.

Educational institutes in Sri Lanka face unique challenges in comparison to developed countries. A strong understanding of these obstacles allows for taking suitable actions for guarantee e-learning success. It is expected that findings of this research will offer a beneficial evidenced based source of information for academics, administrators and decision makers involved in planning, design and implementation of e-learning.

References

[1] DocPlayer, "Chapter 2. Literature review. Integrated in many universities education programs, shifting from traditional way of,". [Online]. Available: http://docplayer.net/14653909-Chapter-2-literature-review-integrated-in-many-universitieseducation-programs-shifting-from-traditional-way-of.html.

[2] D. Garrison, W. Archer and T. Anderson, E-learning in the 21st century, 1st ed. London: Routledge/Falmer, 2002.

[3] S. Harandi, "Effects of e-learning on Students' Motivation", *Procedia - Social and Behavioral Sciences*, vol. 181, pp. 423-430, 2015.

[4] A. Wang and M. Newlin, "Predictors of web-student performance: the role of self-efficacy and reasons for taking an on-line class", Computers in Human Behavior, vol. 18, no. 2, pp. 151-163,2002. [Online].Available:https://www.researchgate.net/publication/222318769_Predictors_of_web-student_performance_The_role_of_self-efficacy_and_reasons_for_taking_an_on-line_class.

[5] Coldwell, Jo, Craig, Annemieke, Paterson, T. and Mustard, Jamie 2008, Online students : relationships between participation, demographics and academic performance, *Electronic journal of e-learning*, vol. 6, no. 1, pp. 19-30.

[6] A. Picciano, "Beyond student perceptions: Issues of interaction, presence, and performance in an online course", Citeseerx.ist.psu.edu, 2017. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.6506.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, Nov. 1996.

[8] Srivastava, R. Cooley, M. Deshpande and P. Tan, "Web usage mining", *SIGKDD Explor. Newsl.*, vol. 1, no. 2, p. 12, 2000.

[9] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann Publishers In, 2006.

[10] leejhon 82, "Data mining vs data collection," in *Data*, Import.io, 2014. [Online]. Available: https://www.import.io/post/data-mining-vs-data-collection/.

[11] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use Moodle courses," *Computer Applications in Engineering Education*, vol. 21, no. 1, pp. 135–146, Jun. 2010.

[12] D. Krpan and S. Stankov, "Educational data mining for grouping students in e-learning system," pp. 207–212, Jan. 2012. [Online]. Available: https://www.researchgate.net/publication/261390213_Educational_Data_Mining_for_Gro upi ng_Students_in_E-learning_System.

[13] "Data mining quick guide," www.tutorialspoint.com, 2016. [Online]. Available: https://www.tutorialspoint.com/data_mining/dm_quick_guide.htm.

[14] A. Bovo, S. Sanchez, O. Héguy, and Y. Duthen, "Analysis of students clustering results based on Moodle log data," Academy Publisher, 2015. [Online]. Available: http://oatao.univtoulouse.fr/13222/.

[15] A. Bovo, S. Sanchez, O. Héguy, and Y. Duthen, "Clustering Moodle data as a tool for profiling students." (2013) In: International Conference on e-Learning and e-Technologies in Education - ICEEE, 2013, 23 September 2013 - 25 September 2013 (Lodz, Poland). [Online]. Available: http://oatao.univ-toulouse.fr/13223/1/bovo_13223.pdf.

[16] D. Krpan and S. Stankov, "Educational data mining for grouping students in e-learning system," pp. 207–212, Jan. 2012. [Online]. Available: https://www.researchgate.net/publication/261390213_Educational_Data_Mining_for_Gro upi ng_Students_in_E-learning_System.

[17] M. A. Hogo, "Evaluation of e-learning systems based on fuzzy clustering models and statistical tools," vol. 37, no. 10, pp. 6891–6903, Oct. 2010. [Online]. Available: https://www.researchgate.net/publication/223087426_Evaluation_of_elearning_systems_based_on_fuzzy_clustering_models_and_statistical_tools.

[18] S. Ventura, "Data mining algorithms to classify students," 2016. [Online]. Available: https://www.academia.edu/2662260/Data_mining_algorithms_to_classify_students.

[19] E. Garcia, C. Romero, S.Ventura, C. de Castro, and T. Calders, "Association rule mining in learning management systems," in *Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*. Informa UK, 2010, pp. 93–106.

[20] B. M. Bidgoli, P. N. Tan, and W. F. Punch, "Mining Interesting Contrast Rules for a Web-based Educational System," in *Computer Science Department, Michigan State University*, 2004. [Online]. Available: https://pdfs.semanticscholar.org/095b/897a18dc501d1cf565f5f38fe82098345575.pdf.

[21] O. R. Zaane and J. Luo, "Towards Evaluating Learners’ Behaviour in a Web-Based Distance Learning Environment," 2001. [Online]. Available: https://webdocs.cs.ualberta.ca/~zaiane/postscript/icalt.pdf.

[22] R. Mazza and L. Botturi, "Graphical interactive student monitoring tool for Moodle," 2007. [Online]. Available: http://gismo.sourceforge.net.

[23] R. Mazza, M. Betton, M. Faré, and L. Mazzola, "MOCLog – Monitoring Online Courses with log data," 1 st Moodle ResearchConference, 2012, pp. 132–139. [Online]. Available:http://research.moodle.net/54/1/17%20-%20Mazza%20-%20MOCLog%20 %20Monitoring%20Online%20Courses%20with%20log%20data.pdf

[24] R. Mazza and V. Dimitrova, "CourseVis: A graphical student monitoring tool for supporting instructors in web-based distance courses," International Journal of HumanComputer Studies, vol. 65, no. 2, pp. 125–139, Feb. 2007.

[25] E. Popescu and D. Cioiu, "EMUSE - integrating web 2.0 tools in a social learning environment," in *Advances in Web-Based Learning - ICWL 2011*. Springer Science + Business Media, 2011, pp. 41–50. [Online]. Available: software.ucv.ro/~epopescu/docs/publications/ICWL_2011.pdf

[26] R. Mazza and V. Dimitrova, "Visualising student tracking data to support instructors in web-based distance education," pp. 154–161, May 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1013393.

[27] [Online].Available:ftp://ftp.software.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CHAID.pdf. Accessed: Mar. 5, 2017.

[28] "CLASSIFICATION METHODS," in *University of Minnesota Duluth*. [Online]. Available: http://www.d.umn.edu/~padhy005/Chapter5.html c

[29] C. Software, "Linear regression model," 2017. [Online]. Available: http://www.camo.com/rt/Resources/linear_regression_model.html.

Appendix

Appendix A: Description of attributes in the data set

Table A.1: Description on modules

Attribute Name	Description	Example values	Special Characteristics
code_presentation	Year with B/J	2013J, 2013B,…	B means course module starts at February and J means course module starts at October.
length	In days	286	

Table A.2: Description of different types of assessments

Attribute Name	Description	Example values	Special Characteristics
id_assessment	Id number for an assessment	14991,…	1. Tutor Marked Assessment – TMA 2. Computer Marked Assessment – CMA 3. Final Exam Assessment scores are stored in StudentAssesment .vle.
assessment_type	Three types of assessments	TMA,CMA,Exam	
date	days for the final submission date considering module start date as 0	215,…	
weight	In %. Exam is cinsider separately as 100%. Other assignment totally 100%	10 20 …..	

Table A.3 Description on involvement of different types of assessments

Attribute Name	Description	Example values	Special Characteristics
date_submitted	The number of days since the start of the module presentation	18, 22,…	
is_banked	Indicating whether the assessment results transferred from the previous attempts.	0 or 1	
score	Student score for the assessment from 0 to 100. Below 40 are fail.	78, 20	

Table A.4: description on interactions on virtual learning system

Attribute Name	Description	Example values	Special Characteristics
Id_site	Id number for the material	546943	
activity_type	Role associated with the model material	resources, autocontent,…	
week_from	Week that material planned to be used	(missing)	
week_to	Week until the material planned to be used	(missing)	

Table A.5: Details of students

Attribute Name	Description	Example values	Special Characteristics
Id_student	Unique id for a student	11391, …	Index of Multiple Deprivation (UK government qualitative study of deprived areas.) This index include Income, Employment,
Gender	Student’s gender	M or F	
Region	Where student live while taking the course	Scotland, Wales, …	
highest_education	Student education level	A level or equivalent, …	
imd_band	Specifies the Index of Multiple Deprevation	20-30%, 30-40%, …	Health Deprivation and Disability, Barriers to Housing and Services, Crime, Living Environment). There are some missing values in IMD band.
age_band	Band of the student’s age	35-55, <=55, …	
num_of_prev_attempts	Number of times student has attempted to a module	0 to 6	
studied_credits	Total number of credits for the modules that a student is currently studying	240, 60, …	
Disability	State whether the student is differently abled.	Y or N	
final_result	Student’s final result	Distinction, Pass, Fail, Withdraw	

Table A.6 Details of registration of students

Attribute Name	Description	Example values	Special Characteristics
date_registration	No. of days from the start of the presentation.	-30, 120 (missing values)	Negative values for the days of registration means number of days before the start of the module presentation.
Date_unregistration	No. of days of getting unregistration. These students' final results are shown as 'Withdraw'.	-196, 12, (missing values)	

Table A.7 Description of activity types used in virtual environment

Activity type	Description
Forum	Teacher provides subject matter, student makes discussions
homepage	Include links for chapters of a module (link color get changed after used refer the module), registration details, payment details, contact information and upcoming events.
oucontent	Represents content of assignments. Seven computer marked assessments where marks are not added for the final result and deadline is before the exam date Five teacher marked assessments where marks are added for the final results and con be submitted online or via post.
quiz	Quizzes with correct answers and explanations, unlimited attempts, marks are not added for the final result and no deadlines
resources	Contains hyperlinks to books, journals and newspaper articles.
subpage	Points to other sites like Youtube, Google play, Open Research Online and FutureLearn.